

# Improving students' and teachers' judgments of student performance in primary school mathematics problem solving



Sophie Oudman

Improving Students' and Teachers' Judgments  
of Student Performance in Primary School  
Mathematics Problem Solving

Vera Sophie Oudman

The research reported in this dissertation was funded by the Dutch Ministry of Education, Culture and Science (grant number OCW/PromoDoc/1065001). It was carried out in the context of the Interuniversity Centre for Educational Sciences (ICO).

**ico**

Cover design: Danaë ten Hoedt

Layout and printing: Proefschrift All In One

ISBN: 978-94-93315-49-5

DOI: <https://doi.org/10.33540/1729>

© 2023 Sophie Oudman

All rights reserved. No part of this dissertation may be reproduced or transmitted in any form or by any means without the prior permission of the author, or when applicable, of the publishers of the scientific papers.

# **Improving Students' and Teachers' Judgments of Student Performance in Primary School Mathematics Problem Solving**

**Verbeteren van de inschattingen die basisschoolleerlingen en  
-leerkrachten maken van de leerlingprestaties op rekentaken**

(met een samenvatting in het Nederlands)

## **Proefschrift**

ter verkrijging van de graad van doctor aan de Universiteit Utrecht  
op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling,  
ingevolge het besluit van het college voor promoties  
in het openbaar te verdedigen op

maandag 15 mei 2023 des middags te 2.15 uur

door

**Vera Sophie Oudman**

geboren op 14 februari 1990  
te Groningen

**Promotor:**

Prof. dr. T.A.J.M. van Gog

**Copromotor:**

Dr. J.E. van de Pol

**Beoordelingscommissie:**

Prof. dr. A.B.H. de Bruin

Prof. dr. P.H.M. Drijvers

Prof. dr. L. Kester

Prof. dr. K. Scheiter

Prof. dr. J.W.F. van Tartwijk

# Contents

<b>Chapter 1</b>	Introduction	6
<b>Chapter 2</b>	Effects of Self-scoring Their Math Problem Solutions on Primary School Students' Monitoring and Regulation	16
<b>Chapter 3</b>	Students' Monitoring and Regulation Accuracy Awareness	46
<b>Chapter 4</b>	Effects of Different Cue Types on the Accuracy of Primary School Teachers' Judgments of Students' Mathematical Understanding	68
<b>Chapter 5</b>	Effects of Cue Availability on Primary School Teachers' Accuracy and Confidence in Their Judgments of Students' Mathematics Performance	94
<b>Chapter 6</b>	Primary School Teachers' Judgments of Their Students' Monitoring and Regulation Skills	128
<b>Chapter 7</b>	Summary and Discussion	160
	Nederlandstalige Samenvatting (Summary in Dutch)	178
	Supplementary Materials	188
	References	224
	Curriculum Vitae	236
	List of Publications	237
	ICO Dissertation Series	240
	Dankwoord (Acknowledgements)	244



# **Chapter 1**

## **Introduction**



In primary school, teachers and students share responsibility for students' learning progress. On the one hand, teachers are expected to provide 'adaptive' or 'differentiated' instruction: making instructional decisions that are adapted to the diverse needs of individual students (Parsons et al., 2018; Tomlinson et al., 2003; Van de Pol et al., 2010). On the other hand, primary school students are increasingly expected to (learn to) self-regulate their learning, as this lays the foundation for students' lifelong learning (Bjork et al., 2013; OECD, 2022) and beneficially influences students' academic achievement (Dent & Koenka, 2016). Consequently, teachers are also increasingly expected to help students develop their self-regulated learning skills (Dignath & Büttner, 2018).

Monitoring and regulation are two central processes in both self-regulated learning (De Bruin & Van Gog, 2012; Griffin et al., 2013) and providing adaptive instruction (Shavelson & Stern, 1981; Thiede et al., 2019; Van de Pol et al., 2011). Monitoring refers to evaluations of individual students' performance. Regulation refers to decisions about what subsequent activities (e.g., restudy, additional practice, or additional instruction) learners should engage in to improve their performance. For those regulation decisions to meet students' actual needs (i.e., be adaptive to their current level of performance), it is critical that monitoring judgments of student's current level of performance are accurate. If students overestimate their own performance, or teachers overestimate their students' performance, the students may quit studying or practicing too early. If students underestimate their own performance (which seems to be rarer; De Bruin et al., 2017; Kruger & Dunning, 1999), or teachers underestimate the students' performance (which, again, seems to be rarer; Urhahne & Wijnia, 2021), the students will spend valuable study time on activities that are already mastered rather than on those they need to learn. In other words, accurate monitoring is a necessary (though not always sufficient) condition for accurate regulation (Dunlosky & Rawson, 2012; Metcalfe & Finn, 2008; Pintrich, 2000; Winne & Hadwin 1998; Zimmerman, 2000).

Thus, for self-regulated learning to be effective, students have to be able to accurately evaluate their own performance and make regulation decisions accordingly (De Bruin & Van Gog, 2012; Griffin et al., 2013). However, primary school students' monitoring and regulation judgments are often inaccurate, for instance when memorizing information (e.g., Roebers et al., 2014), when learning from texts (e.g., De Bruin et al., 2011), and when learning to solve problems (e.g., Baars et al., 2014a; Boekaerts & Rozendaal, 2010; García et al., 2016). Therefore, researchers are looking for ways to help primary school students to improve the accuracy of these judgments. To date, most research on such interventions has been conducted in the context of memorizing information (e.g., Lipko-Speed, 2013; Van Loon et al., 2013; Van Loon & Roebers, 2017) and text comprehension (e.g., De Bruin et al., 2011; Kostons & De Koning, 2017). Relatively little attention has been paid to interventions aimed at improving primary school students' monitoring and regulation judgments when practicing with problem-solving tasks (for an exception see

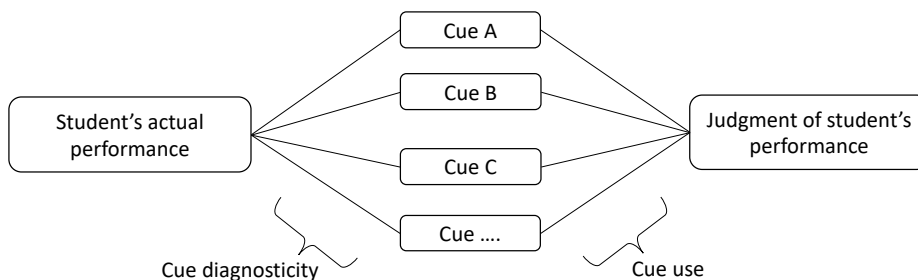
Baars et al., 2014a), even though problem-solving tasks play an important role in primary and secondary school subjects such as mathematics and science.

Similarly, we know that providing students with ‘adaptive’ or ‘differentiated’ instruction—that is, instruction that is adapted to individual students’ needs—promotes students’ learning compared to one-size fits all teaching (Deunk et al., 2018; Parsons et al., 2018; Tomlinson et al., 2003; Van de Pol et al., 2010). However, to be truly adaptive, and thus, optimally effective, teachers need to be able to make accurate monitoring judgments of their students’ performance and make decisions on subsequent activities accordingly (for empirical studies, see Klug et al., 2013; Van de Pol et al., 2011; for review studies, see Thiede et al., 2019; Urhahne & Wijnia, 2021). Yet, prior studies have shown that there is substantial room for improvement in the accuracy of teacher judgments of their students’ performance, for instance when judging their students’ vocabulary, reading comprehension, and mathematical problem-solving skills (for review studies, see Kaufman, 2020; Südkamp et al., 2012; Urhahne & Wijnia, 2021). Therefore, as for students, research is needed on interventions that could improve the accuracy of teachers’ monitoring judgments. At the start of this dissertation research (in 2016), studies on how the accuracy of teachers’ monitoring judgments of their students’ performance can be improved were scarce (for an exception, see Thiede et al., 2015).

In sum, the main aim of the research presented in this dissertation was to gain more insight into how the accuracy of primary school students’ and teachers’ monitoring and regulation judgments of students’ performance on mathematical problem-solving tasks can be improved. To do so, one first needs to look at the origin of inaccuracy in students’ and teachers’ monitoring judgments, which can be explained by the information or *cues* on which they base their judgments.

## **1.1. Improving Student and Teacher Judgments: A Cue-utilization Approach**

As mentioned earlier, both students’ and teachers’ monitoring judgments are often inaccurate, which is a problem as this will also influence accuracy of their regulation judgments. An explanation for the inaccuracy of monitoring judgments lies in the cues on which they base their judgments. Brunswik (1955) used the analogy of a convex lens to explain that humans do not perceive the environment directly, but make inferences, judgments, and decisions through the “lens” of cues (i.e., perceived environmental features). Likewise, research has shown that students (Koriat, 1997) and teachers (Byers & Evans, 1980; Snow, 1968) do not have direct access to the quality of students’ cognitive states, but have to infer, for instance, a students’ current level of performance or degree of understanding from available cues (i.e., pieces of information), see Figure 1.1.

**Figure 1.1***Brunswik's Lens Model Applied to Judgments of Student's Performance*

Cues can be derived from different sources. Students can base their monitoring judgments on cues derived from the to-be-judged task (e.g., complexity or length of the task), experiences during completion of the task (e.g., fluency, response time, or invested mental effort) or general beliefs about themselves (e.g., their interest in the task content or general ability in that school subject; Ackerman, 2019; Koriat, 1997; Thiede et al., 2010). Unlike students, teachers cannot base their judgments on students' experiences during a task, except for what they can observe (e.g., they may be able to infer from students' behavior that they are struggling), but like students, they can base their judgments on task characteristics and their general beliefs about the students. For instance, general student characteristics on which teachers seem to base their judgments of students' academic performance are students' general cognitive abilities (e.g., Kaiser et al., 2015), effort (e.g., Kaiser et al., 2013), disability status (e.g., Hurwitz et al., 2017), or migration background (e.g., Furnari et al., 2017). Such indications of cue use are usually obtained by computing the correlation between the cue and the monitoring judgment, but could also be measured more directly, for instance, by think-aloud protocols (cf. Cooksey et al., 2007).

The reason why monitoring judgments are often inaccurate, is that the available cues differ in the extent to which they are diagnostic (i.e., predictive) of students' actual performance, and students and teachers might not always use highly diagnostic cues. General student characteristics and task characteristics seem to have low diagnosticity for students' performance on a specific task (Ackerman, 2019; Thiede et al., 2010). The diagnosticity of experiential cues, such as fluency, response time, or invested mental effort, seems to be variable and dependent on task characteristics (Ackerman, 2019; Van Gog et al., 2020). The more diagnostic the cues being used, the more accurate a student's or teacher's monitoring judgment will be (Koriat, 1997; Thiede et al., 2010, 2015; Van Loon et al., 2014). Thus, interventions that would give students and teachers access to cues that are more diagnostic of the to-be-judged performance, could improve their monitoring accuracy, and, in turn, their regulation accuracy.

## 1.2. Improving Student Judgments

Research has shown that having students engage in certain activities that require them to demonstrate their understanding of a task, can provide them with access to (experiential) *performance cues* with high diagnosticity: A substantial number of studies on text comprehension amongst primary school, secondary school, and university students have shown that having students generate keywords, summaries, diagrams, drawings, or concept maps after studying a text improved the accuracy of their monitoring judgments (but not necessarily their regulation accuracy; for review studies, see Griffin et al., 2019; Van de Pol et al., 2020). Students' experiences while being involved in such *generative activities* (Fiorella & Mayer, 2016) will give them insight into how well they understood the concepts and (causal) relations described in the text, which is a more diagnostic cue for future test performance than their memory of more superficial details from the text or the fluency they experienced while reading the text.

When learning to solve problems by studying worked examples, students' monitoring accuracy was also found to improve from generative activities. For instance, asking students to complete steps in partially worked-out examples improved monitoring (but not regulation) accuracy of secondary school students (Baars et al., 2013), and asking them to generate the solution to isomorphic practice problems after studying a worked example improved monitoring accuracy of both primary (Baars et al., 2014a) and secondary (Baars et al., 2017) school students. However, it only improved regulation accuracy of the latter.

Another activity that can help students gauge their current level of performance—and thus provides highly diagnostic performance cues—is self-scoring: providing them with a standard of the correct answers to which to compare their own answers. Prior studies on self-scoring were mainly conducted in the field of concept learning and showed that self-scoring improved the monitoring accuracy of primary school students (e.g., Van Loon & Roebers, 2017), adolescents (e.g., Lipko et al., 2009), and adults (e.g., Rawson & Dunlosky, 2007). Van Loon and Roebers (2017) also found that self-scoring improved primary school students' regulation accuracy when learning concepts, but that their regulation judgments were still substantially over-optimistic after self-scoring.

Baars et al. (2014b) investigated the use of standards in problem-solving tasks. They provided secondary school students with the correct worked-out solution procedure against which they could compare their own solutions of biology problems in the domain of heredity and found that students' monitoring accuracy improved (but not their regulation accuracy). As self-scoring is often used in mathematics education in primary schools, this is a good candidate for a potential intervention, and therefore, *the first aim of the research presented in this dissertation was to investigate whether self-scoring would improve primary school students' monitoring and regulation accuracy when learning to solve problems.*

Only improving students' monitoring and regulation judgment accuracy might not be sufficient for improving the effectiveness of students' self-regulated learning behavior, however. Students presumably also need to feel confident about the accuracy of their judgments to act upon them (as suggested by Gabriele et al., 2016; Patterson et al., 2001), which is what we want them to do when their monitoring and regulation judgments are accurate. In contrast, when students make inaccurate judgments, it can be helpful if they feel less confident about the accuracy of their judgments, as acting upon those would hamper their learning. Therefore, research has also started to investigate students' feeling of confidence in their monitoring accuracy. These ratings are also known as second-order judgments (SOJs; Dunlosky et al., 2005; Fritzsche et al., 2018; Händel & Dresel, 2018; Nederhand et al., 2021). When students feel relatively more confident (i.e., providing higher SOJs) about the accuracy of more accurate judgments than of less accurate judgments and vice versa, they show *accuracy awareness* (e.g., Fritzsche et al., 2018; Händel & Dresel, 2018; Nederhand et al., 2021). The few studies available, however, only addressed monitoring accuracy awareness—not regulation accuracy awareness, which is arguably as or even more important for actual study behavior—and were focused on adolescents and young adults, not on primary school children. *Therefore, the second aim of the research described in this dissertation was to explore whether primary school students would be aware of their monitoring and regulation (in)accuracy and whether and how self-scoring would affect their accuracy awareness.*

### 1.3. Improving Teacher Judgments

Interestingly, the activities that seem to provide students with more diagnostic performance cues yield products (e.g., the generated summaries or diagrams, students' answers to prior practice problems, or self-scored tasks) that we might also expect to provide teachers with more diagnostic cues for judging their students' performance. Research on how teachers can be stimulated to focus on more diagnostic performance cues was scarce when I started with this dissertation research. A study by Thiede et al. (2015) provided some evidence that focusing on performance cues improves the accuracy of teacher judgments of students' mathematical performance. They investigated the effects of a professional development program that stimulated primary school teachers to focus more on students' actual understanding and thinking during teaching (e.g., by asking students to articulate their way of reasoning) instead of mainly on the content taught. Teachers who took part in the professional development program indeed made more accurate monitoring judgments of students' mathematical performance than teachers who did not participate in the training program. However, as the authors did not provide information on the cues that teachers presumably used, it remains unclear whether it was indeed the increased focus on highly diagnostic cues, or some other

aspect(s) of the 45-h training that caused the improvement in teachers' monitoring accuracy (e.g., improved mathematical content knowledge). *Therefore, the third aim of the research described in this dissertation was to investigate whether providing primary school teachers with student products from which they can infer performance cues would improve the monitoring accuracy of their judgments of students' performance.*

In parallel to students' self-regulated learning, it also goes for teachers that in order to make effective instructional decisions, they not only need to make accurate judgments of their students' performance, but they also need to be aware of their judgment (in)accuracy (Gabriele et al., 2016). When teachers are aware of their (in)accuracy it is more likely that they make appropriate instructional decisions, either based on judgments that were accurate and in which they have confidence, or by obtaining more information on the judgments that were inaccurate and in which they have less confidence. However, teachers' awareness of their judgment (in)accuracy had—to the best of my knowledge—not yet been investigated. *Therefore, the fourth aim of this dissertation research was to explore to what extent teachers are aware of their monitoring (in)accuracy and whether and how the availability of performance cues affects teachers' awareness of their (in)accuracy.*

Teachers also face another important task. Besides providing students with instruction adapted to their level of performance, teachers also need to help students develop their self-monitoring and self-regulating skills. To do so, teachers need to be able to identify which students might need support; that is, they need to be able to accurately judge the accuracy of their students' monitoring and regulation skills and identify those students who are (highly) inaccurate. This would allow teachers to instruct those students who need it on how to evaluate their performance, on how to make appropriate subsequent decisions, and on when and how to seek help (Azevedo et al., 2008; Dignath & Büttner, 2018). However, the question of whether primary school teachers have accurate insights into how well their students can monitor and regulate their learning had not yet been addressed in prior research. Thus, it is unclear whether—and if so, how—teachers should be supported in making such judgments to improve their ability to support students' development of self-regulated learning skills. Therefore, the fifth aim of the research presented in this dissertation was to explore how well primary school teachers can judge their students' monitoring and regulation skills and what factors influence this judgment process.

## 1.4. Overview of This Dissertation

The five aims outlined above were addressed in five empirical studies, presented in chapters 2 to 6, that were conducted in Dutch grade 6, with 9- to 10-year-old students and their teachers. The data for chapters 2, 3, 5, and 6 were gathered in one large data collection; chapter 4 concerned a separate data collection.

The studies presented in **Chapter 2 and 3** focused on improving students' (awareness of their) monitoring and regulation accuracy. The study described in **Chapter 2** tested whether self-scoring would have beneficial effects on primary school students' monitoring and regulation accuracy when solving procedural mathematics problems; more specifically, multiplication and division problems. Furthermore, it was investigated whether the effects of self-scoring on (potential) improvements in monitoring and regulation accuracy would differ between low- and high-performing students, as this would call for a differential focus in interventions. The study presented in **Chapter 3** explored whether students were aware of their monitoring and regulation (in)accuracy and whether and how this was affected by self-scoring.

The studies presented in **Chapter 4, 5, and 6** focused on improving teachers' (awareness of their) monitoring judgments of students' performance, as well as on their judgments of students' monitoring and regulation accuracy. In the studies reported in **Chapter 4 and 5**, the availability of student cues (i.e., general student characteristics such as their interest in mathematics or nationality) and performance cues (i.e., students' answers or scores on prior practice tasks that presumably would provide highly diagnostic cues) was manipulated. It was investigated whether the availability of performance cues would affect teachers' cue use and improve their monitoring accuracy.

The study presented in **Chapter 4** was concerned with teacher judgments of their students' conceptual understanding of decimal magnitude. The students completed a practice task and follow-up task, while teachers only made judgments about the students' performances on the follow-up task. All teachers made these judgments under three conditions, while having access to: (1) only students' names (i.e., only student cues available), (2) only anonymized students' answers on decimal magnitude practice problems (i.e., only performance cues available), and (3) both students' names and their answers (i.e., both student and performance cues available). Knowing the students' names would give teachers access to student cues, based on their knowledge of the student. These student cues were expected to have low diagnosticity for students' performance on a specific task, in this case: decimal magnitude problems. Having the students' answers on the practice problems available, would allow teachers to infer students' decimal (mis)conceptions—cues that were expected to have high diagnosticity—by analyzing error patterns in students' answers. The teachers were asked to think aloud while making judgments, to measure their cue use.

Identifying the decimal (mis)conceptions from students' answers in the study presented in Chapter 4 would require quite some interpretation by the teachers. Having products available that do not ask for complex interpretations, might make it easier for teachers to use more diagnostic performance cues and *ignore* non-diagnostic student cues, which should improve the accuracy of their monitoring judgments. The study described in **Chapter 5** tested this hypothesis, by providing teachers with performance

cues that do not ask for interpretation: students' scores on prior tasks. All teachers made judgments of how their students would perform on a multiplication and division task, under two conditions: while having access to (1) only students' names (i.e., only student cues available) and (2) both students' names and their scores on similar multiplication and division tasks, completed one week earlier (i.e., student and performance cues available). An indication of teachers' cue use was obtained by computing correlations between teacher judgments of their students' performance and teachers' perceptions (measured by a questionnaire) of general student characteristics (e.g., interest in mathematics, nationality). It was also explored to what extent teachers would be aware of their monitoring (in)accuracy and whether and how this was affected by the availability of performance cues.

The study presented in **Chapter 6** explored how well teachers would be able to judge the accuracy of their students' monitoring and regulation skills. Teachers were not only asked to judge their students' performance and needs but were also asked to indicate what monitoring and regulation judgments they thought their students had made. From these judgments, variables were computed that indicated (the accuracy of) teacher judgments of students' monitoring and regulation accuracy. Moreover, it was explored whether and how (the accuracy of) teacher judgments of students' monitoring and regulation accuracy would be affected by teachers' perceptions of general student characteristics, to gain more insight into how (in)accurate teacher judgments of student monitoring and regulation skills are established.

**Chapter 7** provides a summary and general discussion of the findings from the studies presented in Chapters 2 to 6. Moreover, theoretical and practical implications are addressed, together with suggestions for further research.





# Chapter 2

## Effects of Self-Scoring Their Math Problem Solutions on Primary School Students' Monitoring and Regulation

This chapter was published as: Oudman, S., Van de Pol, J., & Van Gog, T. (2022). Effects of self-scoring their math problem solutions on primary school students' monitoring and regulation. *Metacognition and Learning*, 17(1), 213-239. <https://doi.org/10.1007/s11409-021-09281-9>

Author contributions: All authors contributed to the design of the study. SO recruited participants, collected and analyzed the data, and drafted the manuscript. All authors contributed to critical revision of the manuscript. JvdP and TvG supervised the study.

## Abstract

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

## 2.1. Introduction

Preparing students to become self-regulated learners has become an important goal of primary education. Not only because students are increasingly required to self-regulate their own learning throughout their entire lifetime, for which primary education lays the foundation, but also because of the beneficial effects of self-regulated learning skills on academic achievement (McClelland & Cameron, 2011). Several models exist that describe the phases and cognitive processes that are involved in self-regulated learning somewhat differently (e.g., Pintrich, 2000; Winne & Hadwin 1998; Zimmerman, 2000). However, these models share general features including that three phases can be distinguished in self-regulated learning—a forethought, performance and reflection phase—between which learners switch whenever necessary (Panadero, 2017). Two central processes in most models of self-regulated learning, and in switching between the processes, are self-monitoring (evaluating one's own performance) and self-regulation (controlling one's own study activities; De Bruin & Van Gog, 2012; Panadero, 2017; Griffin et al., 2013). Unfortunately, primary school students' self-monitoring and self-regulation are often inaccurate, and researchers are looking for ways to help them improve these processes (e.g., Baars et al., 2014a; García et al., 2016; Van Loon & Roebbers, 2017). However, relatively little attention has been paid to self-monitoring, and especially self-regulation, when practicing with problem-solving tasks (Van Gog et al., 2020), even though problem solving plays an important role in many primary and secondary school subjects such as mathematics and science (for exceptions, see Baars et al., 2014a, 2018; Boekaerts & Rozendaal, 2010; García et al., 2016; Rutherford, 2017). To be able to optimally support primary school students' self-monitoring and self-regulation when acquiring problem-solving skills, we need to gain a better understanding of their self-monitoring and self-regulation accuracy and what interventions are successful for improving accuracy. Moreover, the differences in accuracy between different student groups (e.g., low- vs. high-performing students) are of interest, as these may call for a differential focus in interventions. Therefore, the present study aims to make a novel contribution to the literature by investigating (1) how students' monitoring accuracy *and* regulation accuracy when practicing problem solving, differs between low- and high-performing students, (2) whether self-scoring, an intervention that has been shown to be effective on other types of learning tasks (Van Loon & Roebbers, 2017), has beneficial effects on students' monitoring and regulation accuracy on problem-solving tasks, (3) whether there is a differential effect of self-scoring for low- and high-performing students, and (4) whether monitoring and regulation are related. Before specifying our research questions in more detail, we first elaborate on prior literature with regard to monitoring and regulation judgments, the differences between low- and high-performing students, and the effects of self-scoring.

### 2.1.1. Monitoring and Regulation Judgments

Accurate monitoring and regulation are necessary for effective self-regulated learning (Dunlosky & Rawson, 2012). In the present study we defined monitoring accuracy as the degree to which students know how well they performed on a task, expressed by the absolute difference between students' judgments of how many problems they answered correctly and the number of problems they actually answered correctly (cf. Baars et al., 2014a; Dunlosky & Rawson, 2012). Across studies, the average accuracy of primary school students' monitoring judgments varies enormously depending on the type of task (Boekaerts & Rozendaal, 2010; Rutherford, 2017) and students' age (accuracy is higher for older children; Destan & Roebbers, 2015; Roebbers et al., 2014). The present study focuses on self-regulated learning of problem-solving tasks in upper primary school, more specifically on 9- to 10-year-old students, practicing computational tasks. To the best of our knowledge, only four studies have focused on primary school students' monitoring judgments of problem-solving tasks. Baars et al. (2014a) studied the effect of self-testing after studying worked examples of water jug problems (subtracting and adding volumes) on 10- to 11-year old students' monitoring accuracy. Rutherford (2017) studied how 7- to 11-year-old students' monitoring accuracy affected their performance on math problems. García et al. (2016) studied 10- to 12-year-old students' monitoring judgments on math problems in relation to online measures of students' metacognitive processes during problem solving. Boekaerts and Rozendaal (2010) studied the effects of math problem type, instruction method, judgment timing, and gender on 10- to 11-year-old students' monitoring accuracy. In all four studies, students mostly overestimated their performance, which is a widespread phenomenon in general (e.g., De Bruin et al., 2017; Dunning & Kruger, 1999).

Self-regulated learning theories generally assume that students' monitoring judgments influence their regulation judgments and that accurate monitoring is a necessary (though not sufficient) precondition for accurate regulation (Metcalfe & Finn, 2008; Pintrich, 2000; Rawson & Dunlosky, 2012; Winne & Hadwin 1998; Zimmerman, 2000). Regulation judgments are decisions on what subsequent learning actions should be taken to reach a learning goal, such as help-seeking or restudying the learning material (Zimmerman, 2000). As regulation judgments directly influence whether and how students continue learning and indirectly influence whether they will master the learning goals or not, making accurate regulation judgments is important. We use the concept "regulation accuracy" to indicate the extent to which the regulation *judgments* are in line with students' *actual need* for regulation, as indicated by experts (thus actual regulation actions were not measured, which is in line with prior studies; e.g., Baars et al., 2014a; Van Loon & Roebbers, 2017). Making accurate regulation judgments appears to be a skill that is strongly under development during the upper years of primary school: Studies about primary school students practicing recall (of word-pairs or information from a

video) showed that regulation judgments of 10- to 11-year old students are influenced more strongly by their monitoring judgments than those of 7- to 8-year-old students and were also far more accurate (i.e., unknown words or definitions were more often selected for restudy). The regulation judgments of the 7- to 8-year-old students seemed to be rather random (Dufresne & Kobasigawa, 1989; Metcalfe & Finn, 2013; Roebers et al., 2014). It remains an open question to what extent primary school students are capable of making accurate regulation judgments in the context of problem solving and whether their regulation judgments are based on their monitoring judgments. If students' regulation decisions are based on their monitoring judgments, then their regulation judgments are presumably too optimistic (given that students' mostly overestimate their performance; Baars et al., 2014a; Boekaerts & Rozendaal, 2010; García et al., 2016; Rutherford, 2017). A potential consequence of regulation that is too optimistic is that students might not seek additional instruction or quit practicing too early and therefore learn less than students who make more accurate judgments.

Prior studies on students' regulation accuracy are mostly in the field of word-pair learning, concept learning, and text comprehension; relatively few have focused on problem solving (e.g., in primary education: Baars et al., 2014a; in secondary education: Baars et al., 2013, 2017; Kostons et al., 2012; for a review of these studies see Van Gog et al., 2020). In these studies, regulation judgments involved students being asked to select word pairs, definitions, texts or worked examples for *restudy* (e.g., Baars et al., 2014a, 2017, 2013; Dunlosky & Rawson, 2012; Metcalfe & Finn, 2008; Van de Pol et al., 2019; Van Loon & Roebers, 2017), *allocate study time* to word pairs (e.g., Dufresne & Kobasigawa, 1989), or *select the complexity* of the subsequent problem-solving task (e.g., Kostons et al., 2012). However, common regulatory actions for problem solving in (Dutch) primary school are somewhat different (cf. three most used mathematics lesson books in the Netherlands [Baak et al., 2018; Borghouts et al., 2019a, 2019b] and EDI, a widely applied teaching model: Hollingsworth & Yabarra, 2018). When students have not yet mastered specific problem-solving skills two regulatory actions are most common: (1) Students receive or ask for additional instruction (by the teacher or another student) when they do not understand how to solve the problems, or (2) Students receive or decide to complete additional (comparable) practice problems when they understand how to solve the problems, but still need a relatively long time to solve the problems. When students master a certain type of problem, they can continue working on another/subsequent learning goal. In line with this practice, we defined self-regulation judgments in the present study as students' indications of what they would need: additional instruction, additional practice, both, or nothing.

### 2.1.2. Unskilled and Unaware

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

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

So far, most research on the unskilled-and-unaware effect has focussed on university students and on learning of word pairs or text comprehension. However, there is one study that suggests that this effect also applies to primary school students who are engaged in problem-solving tasks (García et al., 2016); a finding we aimed to replicate in the present study. Moreover, García et al. (2016) suggest that high-performing students may also make more accurate *regulation* judgments (as opposed to monitoring

judgments) than low-performing students, but they did not provide empirical evidence for this suggestion. Because regulation judgments of upper primary school students are not necessarily based on their monitoring judgments (see section 2.1.1), the unskilled-and-unaware effect as it occurs in students' monitoring judgments, may not necessarily be translated into their regulation judgments. For designing interventions aimed at improving students' regulation judgments, it is relevant to know whether low-performing primary school students indeed differ from their high-performing peers in their regulation judgment accuracy and hence need a different kind of intervention or teacher support.

### 2.1.3. Improving Judgment Accuracy: Effects of Self-Scoring

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

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

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



### 2.1.4. The Present Study

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

- RQ1:* Does the unskilled-and-unaware effect apply to primary school students who are involved in problem-solving tasks?
- a. We expected low-performing students to make less accurate monitoring judgments than high-performing students (cf. García et al., 2016).
  - b. We explored whether low-performing students make less accurate regulation judgments than high-performing students.
- RQ2:* How does self-scoring affect students' monitoring and regulation accuracy?
- a. We expected students' monitoring judgments to be more accurate (i.e., almost perfectly accurate) after self-scoring than before self-scoring (cf. Van Loon & Roebers, 2017).
  - b. We expected students' regulation judgments to be more accurate after self-scoring than before self-scoring (cf. Van Loon & Roebers, 2017). We did not necessarily expect these to become near-perfectly accurate, as hindsight bias may play a role here.
- RQ3:* Does the unskilled-and-unaware effect remain after self-scoring? We explored how low- and high-performing students differ in their:
- a. monitoring accuracy after self-scoring.
  - b. regulation accuracy after self-scoring.
- RQ4:* Are students' regulation judgments related to their monitoring judgments? We explored whether students' regulation judgments are related to their:
- a. monitoring judgments before self-scoring, and whether this differs between low- and high-performing students.
  - b. monitoring judgments after self-scoring, and whether this differs between low- and high-performing students.

## 2.2. Method

### 2.2.1. Participants

Thirty-four Dutch sixth-grade classes participated (Dutch sixth grade is similar to US fourth grade in terms of age, i.e., 9- to 10-year-old students).<sup>1</sup> Of the 777 students who attended the 34 participating classes, data from 495 students were included in the analyses of the multiplication tasks and 359 in the analyses of the division tasks. Two hundred ninety students were included in the analyses for both the division and the multiplication task. The students participated in the current study on two different days with one week in between, working on parallel versions of the tasks.<sup>2</sup> Figure 2.1 displays demographics and for which reasons (and how many) students had to be excluded. As Figure 2.1 shows, a substantial number of students was excluded because they: 1) did not answer any problem on one or both days, 2) did not correctly answer any of the problems on both days, or 3) correctly answered all problems on both days. The reason these students were excluded from the analyses is that making accurate judgments would be relatively easy for them, because the tasks are presumably far too complex (1 and 2) or far too easy (3) for these students. Including these students could have distorted the results. To draw meaningful conclusions about monitoring and regulation accuracy of low- and high-performing students, the tasks should be at a suitable level of complexity and not far beyond their reach or far too easy (Kostons et al., 2012).

The data of this study are openly available in an online depository at [https://osf.io/b4rkf/?view\\_only=70db7d9dff84f3583a22d64e276a5f1](https://osf.io/b4rkf/?view_only=70db7d9dff84f3583a22d64e276a5f1).

### 2.2.2. Materials and Measures

#### 2.2.2.1. Student Performance

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

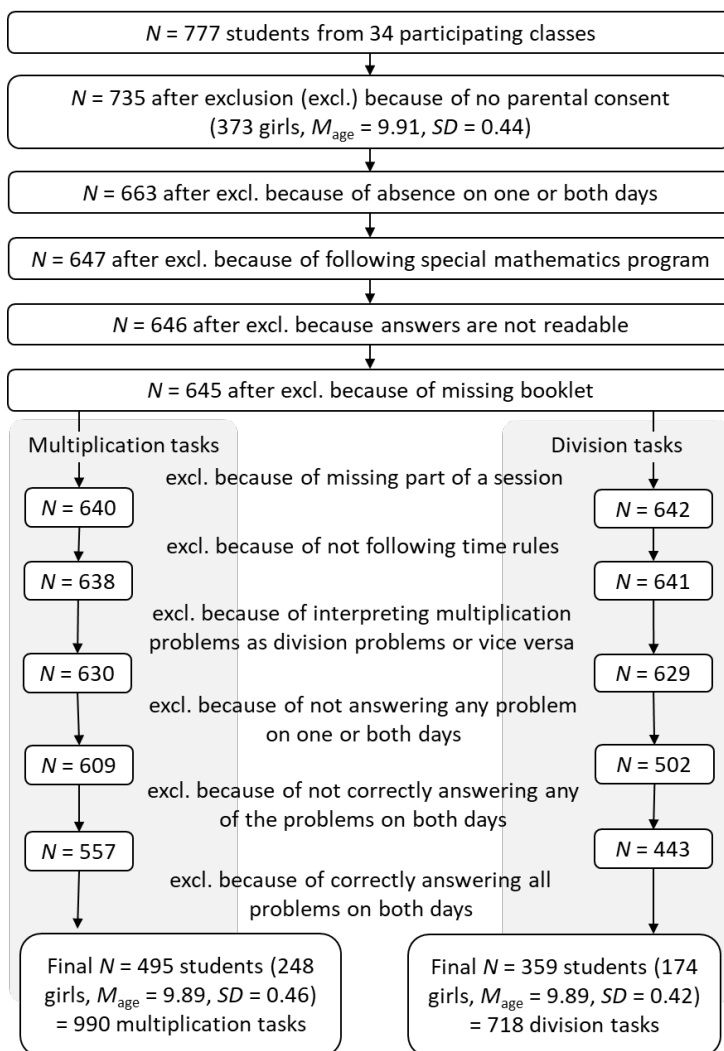
---

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

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

**Figure 2.1**

*Flowchart of Why and How Many Students Were Excluded From All Analyses*



Note. Multivariate outliers were defined for each analysis separately and are still included in the numbers in this flowchart.

### 2.2.2.2. Monitoring Judgment (Accuracy)

After students completed the multiplication or division task they answered the question “How many of the 6 multiplication/division problems do you think you solved correctly?” on a 7-point scale ranging from 0 to 6 (i.e., monitoring judgment before self-scoring). After self-scoring a set of problems, students answered the question “How many of the

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

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

### **2.2.2.3. Regulation Judgment (Accuracy)**

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

To determine the accuracy of students’ regulation judgments, we first coded students’ *actual need for intervention*, based on a coding scheme we developed. We considered students to be in need of an additional intervention when they made (1) procedural errors, which could consist of using a wrong strategy or making wrong use of a correct strategy (these errors are described by Van Zanten et al., 2007), (2) computational errors, indicating sloppiness or a lack of fluency with basic math facts (Calhoon et al., 2007), or (3) exceeding the time limit of 10 minutes (which, based on the opinion of two math

experts and three experienced sixth-grade teachers is the maximum amount of time students who have automated the procedures would need), indicating that students did not yet automatize the procedures or, again, lack fluency with basic math facts. Examples of procedural and computational errors are described in Table S2.1, in section 2.1 of the Supplementary Materials. We had insight into how students performed the computations, because they had been instructed to use space within the booklets as scrap paper and write out their computations. Students' tasks could not be coded item by item, because procedural errors could only be recognized as such when students made the same error multiple times. Therefore, students' needs were defined at the task level. We distinguished four categories. First, students who correctly answered five or six out of six problems within 10 minutes were considered to *not need additional instruction or practice*. Second, students who correctly answered five or six problems in more than 10 minutes or had more than one incorrect answer caused by computational errors were considered to *need additional practice*. Third, students who made procedural errors (specifically, students who gave more than one incorrect answer caused by the use of a wrong strategy or more than two incorrect answers caused by the wrong use of a correct strategy) were considered to *need additional instruction (and practice afterwards)*. We combined the needs "additional instruction" and "additional instruction and practice" into one, because we were not able to decide which of the two needs was more appropriate based on students' work (i.e., their answers and computations that were written out on the scrap paper). In (Dutch) classroom practice, teachers commonly decide during additional instruction to what extent a student needs additional practice afterwards, based on students' understanding during the additional instruction (cf. Baak et al., 2018; Borghouts et al., 2019a, 2019b; Van de Pol et al., 2010). Because actually giving additional instruction was not part of the procedure of our study, we did not know whether or not additional practice after instruction would be needed. However, it is arguably most important that students recognize their need for additional instruction, regardless of whether additional practice would then follow or not (because this can still be decided by the teacher during the additional instruction). Thus, when students' performance indicated they needed additional instruction (and perhaps practice), the researchers scored both the student judgment "additional instruction" and the judgment "additional instruction and practice" as being accurate. Fourth, students who made one procedural error *and* gave one or more incorrect answers caused by computational errors, were considered to *need additional instruction (and practice afterwards) or additional practice only* (in other words, we did not know which intervention was most applicable to the student). When this double code was assigned by the researchers the student judgments "additional instruction", "additional instruction and practice" and "additional practice" were scored as accurate. The detailed coding scheme is depicted in Figure S2.1, in section 2.1 of the Supplementary Materials. To check the interrater reliability of the

coding scheme, two coders (the first author and a research assistant) independently coded 10% of the 409 multiplication and 201 division tasks that could not be coded by preprogrammed rules (see Figure S2.1. for these rules). The interrater reliability was substantial for the multiplication tasks ( $\kappa = .70$ ) and almost perfect for the division tasks ( $\kappa = .85$ ; Landis & Koch, 1977). In case of disagreement, the coders reached consensus through discussion. The first author coded the other 90% of the tasks.

Students' *regulation bias* was measured by comparing their regulation judgment to their regulation need (determined by the researchers), which resulted in values ranging from -2 to +2, with values below zero indicating overestimation of their need for intervention and values above zero indicating underestimation of their need for intervention (see Table S2.2, in section 2.2 of the Supplementary Materials). Again, regulation bias is only reported in the descriptive results, but not analyzed. We used students' *absolute regulation accuracy* for the analyses, which is the absolute value of students' regulation bias. It ranged from 0 to 2, with values closer to zero indicating more accurate monitoring judgments.

### 2.2.3. Procedure

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

3 Hence, waiting time differed across students. Waiting time did not significantly relate to students' monitoring or regulation accuracy before self-scoring ( $p_{\text{monitoring\_multiplication}} = .051$ ,  $p_{\text{regulation\_multiplication}} = .574$ ,  $p_{\text{monitoring\_division}} = .635$ , and  $p_{\text{regulation\_division}} = .105$ )

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

answers aloud). The following monitoring judgment, regulation judgment and second-order regulation judgment were again read aloud by the experimenter. This procedure of completing the second booklet was then repeated for the division task. This entire procedure (but with isomorphic problems) was repeated exactly one week later.

## 2.2.4. Analyses

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

All analyses were performed separately for the multiplication and the division task. We defined four levels in our data: self-scoring condition (before/after; level 1), day (level 2), student (level 3), and class (level 4). We performed multilevel regression analyses in Mplus version 8 (Muthén & Muthén, 1998-2017), using maximum likelihood estimation with robust standard errors (MLR) which is robust to non-normality.<sup>5</sup> The class level was modeled by use of the “Complex” function, because we were not interested in the (fixed or random) effects on this level, we only wanted to account for the non-independence of observations within classes. For the research questions 1A, 1B, 3A, and 3B, about the unskilled-and-unaware effect, the fixed effects were tested at the student level. This means that students' monitoring and regulation accuracy were averaged across the two days. For research questions 2A and 2B, on the effects of self-scoring on students' monitoring and regulation, the fixed effects were tested at the self-scoring condition (before/after) level. For research questions 4A and 4B, on the relation between monitoring and regulation, the fixed effects were tested at the day level. For each of the research questions 1A, 1B, 3A, and 3B, four models were analyzed: two for the multiplication task (intercept only model and model with predictors) and two for the division task

---

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

(intercept only model and model with predictors; resulting in 16 models for RQ1 and RQ3). For both RQ2A and RQ2B, the four models as described for RQ1 and RQ3 had to be performed three times: for the whole sample and for the low- and high-performing students (resulting in 24 models for RQ2). For both RQ4A and RQ4B six models were estimated; three for the multiplication task (whole sample/ low-performing students/ high-performing students) and three for the division task (whole sample/ low-performing students/ high-performing students). For RQ4A and RQ4B, no intercept only models were analyzed, because nominal variables do not have a measure of variance, resulting in 12 models for RQ4. Thus, in total, 52 models were analyzed, which are all presented in section 2.3 of the Supplementary Materials. Only the results that are relevant for answering the research questions are presented in the Results section.

In each of the 52 multilevel models, zero to 13 cases (a maximum of 5.4% of the data) were identified as multivariate outliers. We were mainly interested in the results of the analyses without outliers to avoid drawing conclusions that are potentially affected by extreme cases in our data. For transparency we additionally ran the analyses also with outliers. When this led to differences in statistical significance of effects (this was the case for none of the fixed effects and for six variance components), we additionally reported the effects of the analyses with outliers in section 2.3 of the Supplementary Materials.

## 2.3. Results

### 2.3.1. The Unskilled-and-Unaware Effect Before Self-Scoring (RQ1)

#### 2.3.1.1. Monitoring Accuracy (RQ1A)

Descriptive statistics are presented in Table 2.1. Low-performing students on average substantially overestimated their performance, especially on the multiplication task ( $M_{\text{multiplication}} = 1.34$ ;  $M_{\text{division}} = 0.58$ ). High-performing students slightly underestimated their performance on the multiplication task ( $M = -0.18$ ), but on the division task overestimation and underestimation cancelled each other out ( $M = -0.10$ , which was not significantly different from 0, see monitoring bias before self-scoring in Table 2.1). For both tasks, skill group was a statistically significant predictor of absolute monitoring accuracy before self-scoring, with high-performing students making more accurate monitoring judgments than low-performing students (Table 2.2).



**Table 2.1**

*Means (M) and Standard Deviations (SD) of the Main Variables of This Study*

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

Note. HP = high-performing. LP = low-performing. Means are across both days.

<sup>a</sup> Values closer to zero indicating more accurate monitoring judgments.

<sup>b</sup> This value does not significantly differ from 0,  $p > .05$ .

**Table 2.2**

*Unstandardized and Standardized Coefficients for the Comparison of Low- Versus High-Performing Students' Absolute Monitoring and Regulation Accuracy, Before and After Self-Scoring*

	Multiplication			Division		
	<i>B</i>	<i>SE</i>	$\beta$	<i>B</i>	<i>SE</i>	$\beta$
Absolute Monitoring Accuracy						
Before self-scoring	-0.84***	0.12	-0.91	-0.30**	0.11	-0.38
After self-scoring	-0.07*	0.03	-0.58	0.02	0.03	0.15
Absolute Regulation Accuracy						
Before self-scoring	-0.26***	0.05	-0.44	-0.03	0.05	-0.06
After self-scoring	-0.23***	0.04	-0.78	-0.02	0.04	-0.04

*Note.* Low-performing was coded as 0, high-performing as 1. The full output of the analyses, including intercepts and random effects, are displayed in Tables S2.3 and S2.7 in section 2.3 of the Supplementary Materials

\*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

### 2.3.1.2. Regulation Accuracy (RQ1B)

Low-performing students underestimated their need for intervention, that is, they thought they needed less additional instruction and practice than they actually needed. High-performing students only slightly overestimated their need for intervention for the division task, but for the multiplication task overestimation and underestimation cancelled each other out (see regulation bias before self-scoring in Table 2.1). Skill group was a statistically significant predictor of absolute regulation accuracy on the multiplication task before self-scoring, with high-performing students making more accurate regulation judgments than low-performing students. On the division task, however, low- and high-performing students did not significantly differ in their absolute regulation accuracy (Table 2.2).

## 2.3.2. Effects of Self-Scoring on Students' Monitoring and Regulation Accuracy (RQ2)

### 2.3.2.1. Monitoring Accuracy (RQ2A)

Table 2.3 shows that the whole sample of students and the subsets of low- and high-performing students made on average more accurate monitoring judgments after self-scoring, compared to before self-scoring, both on the multiplication and division tasks. The increase for the whole sample was on average about one problem on a set of six problems for the multiplication and division task. Monitoring became close to accurate after self-scoring (Table 2.1). However, 11.4% (for multiplication) and 9.3% (for division)

of the students from the whole sample still inaccurately judged their performance even though they had been provided with the correct answers. Note that also high-performing students on average slightly overestimated their performance after self-scoring ( $M = 0.07$  for multiplication;  $M = 0.08$  for division; Table 2.1). We additionally explored the causes for the inaccurate monitoring judgments after self-scoring, which are presented in Table 2.4. The most frequent cause was that students did not accurately self-score their answers.

**Table 2.3**

*Effects of Self-Scoring on Absolute Monitoring/Regulation Accuracy*

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

*Note.* Before self-scoring was coded as 0, after self-scoring as 1. The full output of the analyses, including intercepts and random effects, are displayed in Tables S2.4, S2.5 and S2.6 in section 2.3 of the Supplementary Materials.

\*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

**Table 2.4**

*Percentages of Causes of Inaccurate Absolute Monitoring After Self-Scoring*

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

### 2.3.2.2. Regulation Accuracy (RQ2B)

Across the whole sample, students made more accurate regulation judgments on both tasks after self-scoring, compared to before self-scoring. The regulation accuracy of the subset of low-performing students and high-performing students only increased significantly for multiplication, but not for the division task (Table 2.3). After self-scoring,

the overestimation and underestimations of high-performing students cancelled each other out for both tasks. Low-performing students still underestimated their need for intervention (Table 2.1).

We additionally explored the frequency of regulation judgment errors. Students who are most “in danger” of not effectively self-regulating their learning process are those who think they do not need any further intervention whereas they actually need one (additional practice or instruction). Therefore it would be relevant to know how often this kind of error occurs. Table 2.5 shows that on the multiplication task 25% and on the division task 13% of all students made this specific judgment error before self-scoring. For the whole sample of students and for the subset of low-performing students the frequency of this judgment error decreased substantially after self-scoring, especially for the multiplication task. For the high-performing students the frequency hardly decreased after self-scoring for the multiplication task and they made this judgment error even more after self-scoring than before self-scoring, on the division task. At the same time, at least 75% of all students knew whether or not an intervention was needed (Table 2.5)

**Table 2.5**

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

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

### 2.3.3. The Unskilled-and-Unaware Effect After Self-Scoring (RQ3)

#### 2.3.3.1. Monitoring Accuracy (RQ3A)

Table 2.2 shows that on the multiplication task, high-performing students made significantly more accurate monitoring judgments after self-scoring than low-performing students, although this difference was only 0.07 on a seven-point scale. For the division task, high- and low-performing students' monitoring accuracy after self-scoring did not differ significantly.

#### 2.3.3.2. Regulation Accuracy (RQ3B)

High-performing students made more accurate regulation judgments after self-scoring than low-performing students on the multiplication task, but not on the division task (Table 2.2).

### **2.3.4. Relation Between Monitoring and Regulation Judgments (RQ4)**

#### **2.3.4.1. Before Self-scoring (RQ4A)**

Table 2.6 presents the results of the logistic regression analysis, measuring the effect of students' monitoring judgments on their regulation judgments (i.e., whether students chose for additional practice versus no intervention or for additional instruction [and practice] versus no intervention). Before self-scoring, the magnitude of the monitoring judgments significantly predicted students' regulation judgments after self-scoring. This was the case for the whole sample of students as well as for the subsets of low- and high-performing students, for both the multiplication and division tasks. When students' monitoring judgments increased with one item, the odds of choosing for additional practice or instruction compared to no intervention became roughly two to seven times smaller.<sup>6</sup>

#### **2.3.4.2. After Self-scoring (RQ4B)**

For the whole sample of students and the low- and high-performing students, the monitoring judgments after self-scoring significantly predicted students' regulation judgments after self-scoring, for both tasks. When students' monitoring judgments increased with one item, the odds of choosing for additional practice or instruction compared to no intervention became roughly two to nine times smaller (Table 2.6).

## **2.4. Discussion**

The current study investigated two key components of primary school students' self-regulated learning, self-monitoring and self-regulation (De Bruin & Van Gog, 2012; Panadero, 2017; Griffin et al., 2013), as well as the interrelation between these two processes. Specifically, we investigated whether the unskilled-and-unaware effect, which states that low-performing students tend to overestimate their performance more than high-performing students, also applies to primary school students' monitoring and regulation judgments with regard to math problem solving (RQ1). In addition, we investigated whether self-scoring students' answers would improve their monitoring and regulation accuracy (RQ2), and whether this differed between low- and high-performing students (RQ3). Finally, we investigated whether students' regulation and monitoring judgments were related, before and after self-scoring (RQ4). Table 2.7 presents an overview of findings on each of those questions, which we discuss below.

---

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

**Table 2.6**

*Effect of Monitoring Judgments on Regulation Judgments, Before and After Self-Scoring*

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

*Note.* "Practice" refers to students' regulation judgment = additional practice needed. "Instruction" refers to regulation judgment = additional instruction (and practice) needed. "Nothing" refers to regulation judgment = no intervention needed. The full output of the analyses, including intercepts and confidence intervals of the odds ratios are displayed in Tables S2.8, S2.9, and S2.10 in section 2.3 of the Supplementary Materials.<sup>a</sup> No students within this sample made the "Instruction" judgment, thus this model could not be analyzed.  
 \*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

**Table 2.7***Overview of the Findings of the Present Study*

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

<sup>a</sup> Findings were comparable when performance was added to the analyses as continuous variable (instead of low/high) and the whole sample was included (instead of only low- and high-performing students), see Tables S2.1 and S2.12 in section 2.3 of the Supplementary Materials.

<sup>b</sup> This effect was not significant when performance was added to the analyses as continuous variable and the whole sample was included, see Table S2.12.

<sup>c</sup> For whole sample and low- and high-performing students.

### 2.4.1. Differences Between High- and Low-Performing Students Before Self-Scoring (RQ1)

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

As for monitoring accuracy, in line with our expectations and the findings of García et al. (2016), high-performing students made more accurate monitoring judgments than low-performing students. Interestingly, we also found an unskilled-and-unaware effect in regulation accuracy, with high-performing students making more accurate regulation judgments than

low-performing students, though only on the multiplication task and not on the division task. An explanation for why the unskilled-and-unaware effect was more pronounced on the multiplication task might lie in the fact that knowledge gaps were easier to identify for students on the division task than on the multiplication task. This explanation is supported by the fact that students made roughly twice as many omission errors (i.e., lack of answers) in the division task than in the multiplication task, where they made more commission errors (i.e., incorrect answers; Table S2.13, in section 2.3 of the Supplementary Materials).

Thus, it seems that, when working on multiplication tasks, students with low skill can easily come up with a “strategy,” even if incorrect, by just multiplying “random” digits of the multiplier with the multiplicand (i.e., commission errors) whereas on the division task, students with low skill seem more likely to get stuck and realize they do not know how to solve the problem, leading to an omission error. This finding also underlines that besides primary school students’ monitoring accuracy (Boekaerts & Rozendaal, 2010; Rutherford et al., 2017) also the unskilled-and-unaware effect seems to be influenced by the nature of the learning task, even if tasks are from the same domain (in our case, mathematics). Other explanations for differential findings across the two types of tasks could lie in (1) the fact that the division task was more difficult than the multiplication task (the difference in performance was 0.5 out of six items; Table 2.1), (2) the fact that students are more familiar with multiplication than with division, because in Dutch primary schools learning multiplications starts at the end of third grade (comparable to US first grade), whereas learning divisions starts at the beginning of fifth grade (for the role of familiarity see Fitzsimmons et al., 2020), and (3) the possibility that order effects played a role, because students always first completed the multiplication task with corresponding judgments and then the division task with corresponding judgments.

Overall, we found substantial evidence for the unskilled-and-unaware effect. There are several possible (non-mutually exclusive) explanations for this difference in monitoring and regulation accuracy between high- and low-performing students (see also section 2.1.2). First, high-performing students may make more accurate monitoring judgments, because they have more knowledge of what good performance entails (Kruger & Dunning, 1999). More accurate monitoring judgments amongst the high-performing students compared to low-performing students might also translate into more accurate regulation judgments, as these two judgments are related (see our findings to the fourth research question). Second, because the tasks impose less cognitive load on high-performing than on low-performing students, high-performing students may have more cognitive capacity available for monitoring, which provides them with more information to base their monitoring judgments on, and subsequently their regulation judgments (Van Gog et al., 2011). Third, low-performing students might suffer the most from wishful thinking as low-performing students’ discrepancy between desired and actual performance is larger compared to high-performing students (Serra & DeMarree, 2016). The fourth possibility,



which is not mentioned before, entails that, because of their high performance, there was less room for high-performing students to overestimate their performance and underestimate their need for intervention, compared to low-performing students, which may have enhanced the unskilled-and-unaware effect. Note though, that in the present study, this would have been mitigated at least partly by the fact that we excluded students who answered all problems correctly or incorrectly. Moreover, even if this “statistical” explanation would apply, our finding that high-performing students make more accurate judgments is still the reality in the classroom, as the math problems we used were part of the actual sixth-grade curriculum.

Besides finding that high-performing students made more accurate monitoring and regulation judgments than low-performing students in general, the type of errors these two groups made, and thus the type of intervention they needed, also differed. While low-performing students mostly made procedural errors and therefore needed additional instruction (and practice afterwards) high-performing students mostly made computational errors and therefore needed additional practice (see Table S2.14, in section 2.3 of the Supplementary Materials). This finding indicates that low-performing students do not only need more support when regulating their learning than high-performing students, but also a different kind of support.

#### **2.4.2. The Effects of Self-Scoring on Monitoring and Regulation Accuracy (RQ2)**

The second research question addressed the effectiveness of an intervention to improve monitoring and regulation accuracy: self-scoring of their solutions, based on a standard (i.e., the correct answers). In line with our expectation and prior research with other tasks (concept learning; Van Loon and Roebers 2017), self-scoring improved the average monitoring accuracy of the whole sample of students and of the subsets of low- and high-performing students. Interestingly, approximately 10% of the students still incorrectly monitored their performance (they were almost always too optimistic), even though they had been provided with the correct answers. Inaccurate monitoring after self-scoring appeared to have three different causes (Table 2.4). First, most of these students did not accurately self-score their answers (in almost all cases students indicated that a specific answer was correct although it was not). This may be caused by students' limited ability or motivation to recognize differences between their answers and the objectively correct information (Dunlosky et al., 2005; Rawson & Dunlosky, 2007). Second, some of these students changed their original answers (we could see this because we changed the pen color in the self-scoring phase). Possibly, they tried to protect their (self-)image. Third, some of these students gave an incorrect monitoring judgment due to an unknown reason, possibly because they did not correctly add up the number of correct answers.

Across the whole sample, students made more accurate regulation judgments on

both the multiplication and division task after self-scoring, compared to before self-scoring. Regulation accuracy of the subsets of low- and high-performing students only increased slightly for the multiplication task and not for the division task. The lack of improvement in regulation accuracy on the division task for the low- and high-performing students could be explained by the fact that students' regulation judgments before self-scoring were more accurate than on the multiplication task (because the knowledge gaps were easier to identify, see above). A possible explanation for the small improvement of the regulation judgments after self-scoring in general could be that hindsight bias played a role here (i.e., the tendency of students to think that they master the computations, although they made mistakes; Fischhoff, 1975); students might have attributed their mistakes to computational errors (which could be an accurate judgment) and may therefore have concluded that no additional instruction and practice was needed in order to do better next time, although additional practice is also needed to prevent computational errors. Another explanation for why inaccurate regulation judgments did not improve, or improved only slightly after self-scoring, might be that students' standards of when they need an additional intervention differ from the standards of experts. For instance, students might think they need additional instruction or practice when they correctly answered three or less out of six problems, whereas we have set this standard at four or less correct answers (based on the opinion of experts).

### **2.4.3. Differences Between High- and Low-Performing Students After Self-Scoring (RQ3)**

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

high-performing students might decide more often for additional practice when this is indeed an appropriate decision. In turn, this might lead to even higher performance.

#### **2.4.4. Relation Between Monitoring and Regulation Judgments (RQ4)**

Whereas theories on self-regulated learning generally assume that students' monitoring judgments are (partially) based on their regulation judgments (Pintrich, 2000; Winne & Hadwin 1998; Zimmerman, 2000), findings of prior studies, which were only in the field of information recall, indicate that this relation starts to appear somewhere in the upper primary school years (Dufresne & Kobasigawa, 1989; Metcalfe & Finn, 2013; Roebbers et al., 2014). Our findings showed that 9- to 10-year-old students' monitoring and regulation judgments regarding math problem solving are, at least to some extent, interrelated. This finding indicates that these students might partially base their regulation judgments on their monitoring judgments, both before and after self-scoring. Interventions aimed at improving students' monitoring accuracy might therefore also, to some extent, translate into improved regulation judgments. However, importantly, our results also indicate that improved monitoring judgments after self-scoring do not always translate into improved regulation judgments: We found that monitoring accuracy became much closer to perfect accuracy after self-scoring than regulation accuracy (Table 2.1). Moreover, in many cases the proportions of students who inaccurately indicated that they did not need an intervention, hardly changed in regulation judgments from before to after self-scoring (Table 2.5). Students' regulation judgments thus seemed to be somewhat resistant to change (especially for high-performing students) or did change, but into another inaccurate decision (especially for low-performing students).

#### **2.4.5. Limitations and Future Research**

One limitation of the present study was that a large number of participants had to be excluded, due to several reasons (see section 2.2.1). Note that in regular classroom practice (in the Netherlands), the excluded students would also be those who would get a different task because they are behind or ahead of the lesson aim for the majority of the students (cf. Baak et al., 2018; Borghouts et al., 2019a, 2019b). In future studies, researchers could consider showing the tasks beforehand to the teachers, ask which of their students would normally not get a task of that difficulty, and only exclude these students. Moreover, we still had a sizable sample overall and in the two subsamples of high- and low-performing students. Whereas there was substantial overlap in students included in the multiplication and division task analyses overall, there was only slight overlap within the low- and high-performing subsamples (i.e., students scoring low on division did not necessarily score low on multiplication and vice versa), which could have played a role in finding the unskilled-and-unaware effect for multiplication, but not for the division task.

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

Future studies might consider including item-by-item judgments in addition to whole task judgments when investigating students' monitoring and regulation judgments in the problem-solving context. The whole task judgments we used in this study are more specific than global judgments of one's own general mathematic skills, but somewhat less specific than item-by-item judgments. Making judgments at this intermediate grain size, at which students judge the extent to which they master a specific skill, is regularly requested of students in primary education (see section 2.2.2.2) and can be useful when students reflect on which specific skills ask for an intervention (Hartwig & Dunlosky, 2017). However, primary school students also make item-specific judgments regularly when working on math problems and future studies could consider comparing the self-regulatory processes involved in solving a single problem and in a complete task (note that in a meta-analysis of Südkamp et al. [2012] no effect of judgment grain size on the accuracy of *teachers'* judgments of student performance was found).

Relatively little research on (improving) monitoring and regulation accuracy has focused on problem-solving tasks in primary education so far. Since the unskilled-and-unaware effect and the effect of self-scoring differed across the multiplication and division task, these effects should be more systematically investigated in different

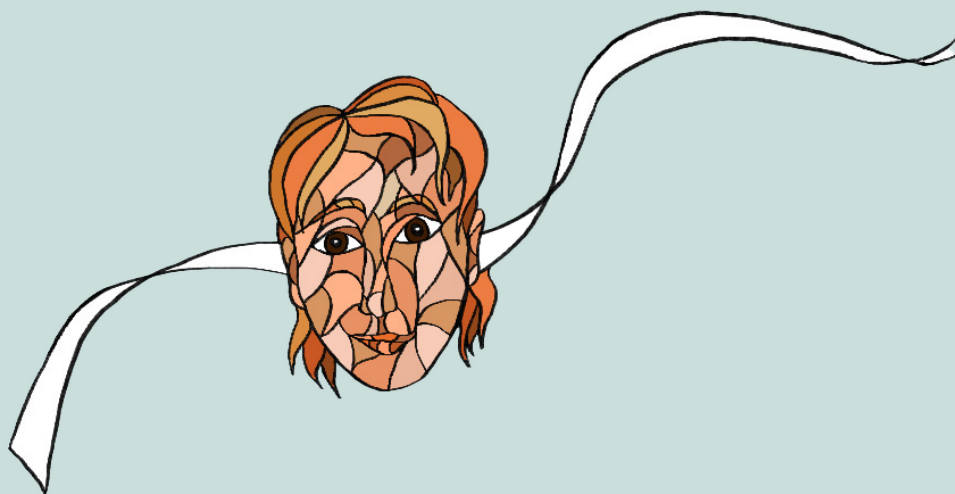
types of problem-solving tasks. For instance, when working on problems that are more ill-structured and more complex than the computational tasks in this study, making accurate judgments and accurately self-scoring one's answers might be more challenging. Moreover, nowadays, schools increasingly start using online learning environments with adaptive math learning programs, in which students receive immediate feedback on their performance. It would be valuable to investigate how different groups of primary school students differ in their help-seeking behavior and how this can be improved when working in these environments (cf. e.g., Roll et al., 2011, who investigated the latter for secondary school students).

Last but not least, our findings may be generalized to schools in which it is common practice (as it is in the Netherlands) that students self-score their answers and are encouraged to take self-regulatory actions such as asking for further instruction or terminating/continuing with practice tasks. For schools in which it is not common practice yet, that would consider implementing self-scoring and subsequent self-regulation, our finding that at least 75% of the students in this study accurately indicated whether or not they needed an additional intervention (Table 2.5) is very promising. However, future research should further investigate and confirm whether similar findings would be obtained in schools or countries where self-scoring and taking self-regulatory actions are not yet common.

#### **2.4.6. Practical Implications and Conclusions**

The current study, together with previous studies (Baars et al., 2014a; Boekaerts & Rozendaal, 2010; García et al., 2016; Rutherford, 2017), showed that primary school students' self-monitoring and self-regulation when practicing with problem solving are not optimal and frequently too optimistic. Our study indicates that having 9- to 10-year-old students self-score their math problem solutions is an effective way to increase their monitoring accuracy, and that this partially translates into improved regulation judgments. Thus, the common practice in many Dutch primary schools to have students self-score their answers (Baak et al., 2018; Borghouts et al., 2019a, 2019b) seems to be good practice. While prior research investigated the unskilled-and-unaware effect with regard to monitoring judgments, our study indicated that this effect also applies to regulation judgments and after self-scoring, at least for one of the two tasks used here. Especially the finding that high-performing students still made more accurate monitoring and regulation judgments after self-scoring for one of the two tasks than low-performing students, suggests that low-performing students need more and different support with self-regulating their learning process than high-performing students, when practicing with problem solving.





# Chapter 3

## Students' Monitoring and Regulation Accuracy Awareness

This chapter is submitted for publication as: Oudman, S., Van de Pol, J., Janssen, E. M., & Van Gog, T. (2022). *Students' Monitoring and Regulation Accuracy Awareness* [Manuscript submitted for Publication].

Author contributions: SO, JvdP, and TvG designed the study. SO recruited participants, collected and analyzed the data, and drafted the manuscript. EJ assisted in analyzing the data. All authors contributed to critical revision of the manuscript. JvdP and TvG supervised the study.



## **Abstract**

We investigated primary school students' awareness of their monitoring and regulation (in)accuracy in mathematics and whether self-scoring their answers improved regulation accuracy awareness (i.e., feeling relatively more confident about the accuracy of more accurate than less accurate judgments and vice versa). Students (9-10 years old) from 34 classes made monitoring and regulation judgments on mathematical tasks, rated their confidence in the accuracy of those judgments, self-scored their work, and again made (confidence) judgments. On average, students showed limited awareness of their monitoring and regulation (in)accuracy prior to self-scoring. Self-scoring seemed to improve students' regulation accuracy awareness overall. Yet this effect was limited for low-performing students and for students whose regulation accuracy decreased or stayed equally inaccurate after self-scoring.

## 3.1. Introduction

Students are increasingly expected to become self-regulated learners (OECD, 2022). Accurate self-monitoring (evaluating one's own performance) and self-regulation (decisions on what subsequent learning actions should be taken to reach a learning goal) are critical for effective self-regulated learning (Dunlosky & Rawson, 2012; Griffin et al., 2013). Prior studies have shown that primary school students engaged in mathematics problem solving often make inaccurate monitoring and regulation judgments (e.g., Baars et al., 2014a; García et al., 2016; Boekaerts & Rozendaal, 2010; Oudman et al., 2022b). Therefore, researchers have been looking for ways to help primary school students improve their monitoring and regulation accuracy, with some success (e.g., Baars et al., 2014a; Van Loon et al., 2017; Oudman et al., 2022b). However, only improving students' monitoring and regulation judgment accuracy might not be sufficient for improving the effectiveness of students' actual self-regulated learning activities. Students presumably also need to feel confident about the accuracy of their judgments to act upon them (suggested by Gabriele et al., 2016; Patterson et al., 2001), which is what we want them to do when their monitoring and regulation judgments are accurate. In contrast, when students make inaccurate judgments, it is helpful if they feel less confident about the accuracy of their judgments (as acting upon those would hamper their learning; see section 3.1.2). Students' ratings of their feeling of confidence in their judgment accuracy are also known as *second-order judgments* (SOJ; Dunlosky et al., 2005; Fritzsche et al., 2018; Händel & Dresel, 2018; Nederhand et al., 2021). When students feel relatively more confident (i.e., providing higher SOJs) about the accuracy of more accurate judgments than of less accurate judgments and vice versa, they show *accuracy awareness* (e.g., Fritzsche et al., 2018; Händel & Dresel, 2018; Nederhand et al., 2021).

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

### 3.1.1. Monitoring and Regulation Accuracy

In the present study we defined students' monitoring accuracy as the degree to which students know how well they performed on a mathematical task, expressed by the absolute difference between students' judgments of how many problems they answered correctly and the number of problems they actually answered correctly (cf. Baars et al., 2014; Dunlosky & Rawson, 2012). We use the concept 'regulation accuracy' to indicate the extent to which students' regulation judgments, meaning their evaluation of their need for additional instruction or practice, are in line with students' actual need for intervention, as indicated by experts (cf. Oudman et al., 2022b).

Students' monitoring judgments influence their regulation judgments and accurate monitoring seems a necessary (though not sufficient) precondition for accurate regulation (Dunlosky & Rawson, 2012; Oudman et al., 2022b). That is, if students overestimate their own performance, they are likely to terminate practicing and move on to another task while they do not yet master the skill and would need additional practice or instruction. If they underestimate their own performance (which seems rarer; De Bruin et al., 2017) they are likely to spend time on activities they already mastered rather than on those they need to learn. Hence, students who make inaccurate judgments may learn less than students who make more accurate judgments. Inaccurate judgments might be less problematic, however, when students are aware of the inaccuracy of their monitoring and regulation judgments.

### 3.1.2. Students' Awareness of Their Judgment (In)Accuracy

Students show accuracy awareness when they indicate that they feel relatively more confident about the accuracy of more accurate monitoring/regulation judgments than about the accuracy of less accurate judgments and vice versa (e.g., Fritzsche et al., 2018; Händel & Dresel, 2018; Nederhand et al., 2021). In other words, students who are aware of their (in)accuracy are able to distinguish between their more and less accurate judgments in terms of their feeling of confidence, whereas students who are not aware of their (in)accuracy are not.

Students' awareness of their judgment (in)accuracy might be an important predictor of how students will self-regulate their learning, because students' feeling of confidence in the accuracy of their judgments might affect whether and how they act upon these judgments (suggested by Gabriele et al., 2016; Händel & Fritzsche, 2016; Patterson et al., 2001). For instance, students who make accurate monitoring and regulation judgments about their learning but do not feel confident about their judgment accuracy (i.e., are not aware of the accuracy) might not act upon their judgments and fail to *actually* regulate their learning accordingly (e.g., they might seek additional instruction whereas only additional practice would have sufficed). Students who make *inaccurate* monitoring and regulation judgments but do feel confident about their judgment accuracy (i.e., are unaware of the inaccuracy)

are likely to act upon their inaccurate judgments, which does not lead to effective self-regulated learning. In contrast, students who make inaccurate monitoring and regulation judgments and feel unconfident about their judgment accuracy (i.e., are aware of the inaccuracy) might ask their teachers for help, which could lead to adjusted—and more accurate—decisions. In sum, it is important to not only investigate students' monitoring and regulation accuracy, but also their accuracy awareness, as this might provide more insight into what is needed to improve students' actual self-regulated learning behavior.

### **3.1.2.1. Prior Research into Students' Awareness of Their Judgment (In) Accuracy**

Previous studies have shown that university students are somewhat aware of their *monitoring* (in)accuracy—in that they felt more confident about more accurate monitoring judgments and vice versa (e.g., Fritzsche et al., 2018; Händel & Dresel, 2018; Nederhand et al., 2021). One study (Nederhand et al., 2021) also investigated secondary school students' awareness of their monitoring (in)accuracy and showed that they were not aware of their monitoring (in)accuracy with regard to exams for French, German, and Mathematics. This might suggest that monitoring accuracy awareness is a metacognitive skill that only develops during adulthood or late adolescence, in which case one would not expect to see it in primary school students. However, it could also be the case that accuracy awareness is related to task performance, and is found in students who perform high on the task but not in students who perform low on a task: Studies with university students (Fritzsche et al., 2018; Händel & Dresel, 2018) found that high-performing students (i.e., defined as students whose task performance fell in the fourth quartile) were aware of their monitoring (in)accuracy and their low-performing peers (i.e., whose task performance fell in the first quartile) were not, but it is unclear from the study by Nederhand et al. whether this could explain the findings with secondary school students.

Moreover, another alternative explanation for why Nederhand et al. (2021) did not find secondary school students to be aware of their monitoring accuracy is a methodological one. We presume, as also argued by Fritzsche et al. (2018), that students only show true accuracy awareness when they are able to distinguish between their more and less accurate judgments in terms their feeling of confidence, which asks for analyzing the effects at the *within*-student level (i.e., based on multiple measurements of accuracy and confidence per student). In contrast, Nederhand et al. (2021) analyzed the data at the *between*-student level (i.e., one measurement of accuracy and confidence per student), which answers a slightly different question, namely: do students, who make a more accurate judgment on a task, feel more confident about the accuracy of this judgment, than students who make a less accurate judgment on that task?<sup>1</sup> Thus, the present study

---

1 See section 3.1. in the Supplementary Materials for a more elaborate explanation of the differences between analyses at the within-student and between-student level.

aimed to explore whether primary school students' monitoring accuracy predicts their feeling of confidence about their monitoring accuracy, at the within-student level, and also in particular when they are low- or high-performing.

To the best of our knowledge, *regulation* accuracy awareness has not been investigated at all thus far. Yet, this might be at least as important, as regulation judgments more directly influence whether and how students continue learning than monitoring judgments. Therefore, we also aimed to explore whether primary school students are aware of their regulation (in)accuracy, and how this differs between low- and high-performing students.

### **3.1.3. Effect of Self-Scoring on Students' Accuracy Awareness**

As students' accuracy awareness might play a role in the effectiveness of self-regulated learning (i.e., in whether or not their monitoring and regulation judgments lead to learning activities that are actually adaptive to their needs), it is relevant to find out whether and how interventions that improve students' monitoring and regulation accuracy, also influence students' accuracy awareness. One intervention that is known to improve students' monitoring and regulation accuracy, is self-scoring: Asking students to compare their own answers to the correct answers (Oudman et al., 2022b; Van Loon & Roebers, 2017). In primary education, students are regularly (and increasingly) asked to self-score their answers before making regulation decisions. In the context of mathematics problem solving in primary school, a prior study showed that after self-scoring, students' monitoring accuracy came close to perfect. However, while students' regulation judgments improved, these still deviated substantially from expert judgments of what (if any) activity the students would need to engage in (Oudman et al., 2022b). To effectively impact students' actual learning behavior, it seems important that if students' regulation accuracy improves from self-scoring, they also feel (more) confident about the accuracy of those judgments, and that if they still make inaccurate regulation judgments after self-scoring, they feel (more) unconfident. Hence, in the present study we investigated students' awareness of their regulation accuracy both before and after self-scoring.

### **3.1.4. Present Study**

The present study addressed three research questions (RQ) in the context of mathematics problem solving in primary school. First, we investigated whether, before self-scoring, students showed awareness of their *monitoring* (in)accuracy (meaning that they feel relatively more confident about the accuracy of more accurate monitoring judgments than of less accurate monitoring judgments and vice versa; RQ1a) and whether and how this differed between low-performing (i.e., in the first quartile) and high-performing (i.e., in the fourth quartile) students (RQ1b). Second, we explored whether, before self-scoring,

students showed awareness of their *regulation* (in)accuracy (RQ2a) and whether and how this differed between low- and high-performing students (RQ2b). Third, we explored whether and how self-scoring affected students' regulation accuracy awareness (RQ3a) and whether and how this differed between low- and high-performing students (RQ3b).

With regard to the first research question, we hypothesized that overall, primary school students would not be aware of their monitoring (in)accuracy prior to self-scoring (Hypothesis 1a). This is based on findings with secondary school students by Nederhand et al. (2021; although they based their conclusions on analyses at the between-student level, whereas we are mainly interested in signs of students' accuracy awareness at the within student level (see section 3.1.2.1), which might lead to different conclusions). Moreover, it could be that accuracy awareness might only be found in high-performing students (Hypothesis 1b). Because there is no prior research about students' *regulation* accuracy awareness, we had no specific hypotheses regarding the second and third research question.

## 3.2. Method

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

### 3.2.1. Design

The students participated on two different days with one week in between, working on parallel versions (i.e., with isomorphic problems that have the same solution procedure and difficulty, but different numbers) of a multiplication and a division task. On both days students made similar (second-order) monitoring and regulation judgments, self-scored their answers, and again made judgments. Similar measures on two days were needed to investigate whether students showed accuracy awareness: When students make more/less accurate judgments on one of both days, do they also feel more/less confident about their accuracy?

### 3.2.2. Participants

In the larger project, 34 Dutch sixth-grade classes participated (Dutch sixth grade is similar to US fourth grade in terms of age, i.e., 9-10 years old). Of the 777 students who attended the 34 classes, data from 495 students were included in the analyses of the multiplication tasks ( $M_{\text{age}} = 9.89$ ,  $SD = 0.46$ , 248 girls) and 359 in the analyses of the division tasks ( $M_{\text{age}} =$

9.89,  $SD = 0.42$ , 174 girls). The students in these two separate datasets partly overlapped: 290 students were included in the analyses for both the division and the multiplication task. The reasons for which (and how many) students and tasks had to be excluded are detailed in our prior study (Oudman et al., 2022b), but it is relevant to note here that a substantial number of tasks was excluded because students (1) did not answer any problem on one or both days, (2) did not correctly answer any of the problems on both days, or (3) correctly answered all problems on both days. The reason these tasks were excluded from the analyses is that making accurate judgments would be relatively easy for these students on these tasks, because the tasks were presumably far too complex (1 and 2) or far too easy (3) for them. Including these data could therefore have distorted the results.

### **3.2.3. Materials and Measures**

#### **3.2.3.1. Student Performance**

On both days, students answered a set of six multiplication problems (single-digit multiplicands multiplied by 3-digit multipliers, e.g.,  $6 \times 472$ ) and a set of six division problems (3-digit dividends divided by single-digit divisors, e.g.,  $282 : 6$ ). Parallel versions—with isomorphic problems that have the same solution procedure and difficulty, but different numbers—of the two mathematical tasks were administered on the two days. Students received one point for each problem that was solved correctly, thus the performance scores ranged between 0 and 6 per task. The internal consistencies of the performance scores, in terms of Cronbach's alpha, were in the acceptable to good range (multiplication: .72 and .68; division: .78 and .81 on the first and second day, respectively).

#### **3.2.3.2. Monitoring Judgment Accuracy**

After students completed the multiplication or division task they made a monitoring judgment by answering the question "How many of the 6 multiplication/division problems do you think you solved correctly?" on a 7-point scale ranging from 0 to 6. Students also made a monitoring judgment after self-scoring, but this judgment was not used in the analyses (see section 3.2.3.4).

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

#### **3.2.3.3. Regulation Judgment Accuracy**

Students also made a regulation judgment before and after self-scoring, by indicating which of the following choices was most applicable to them: additional instruction, additional practice, additional instruction and practice, or no additional instruction and no additional practice on the type of problems they just completed. The researchers made

it clear to the students that they would not actually receive the additional intervention. Students' regulation judgments were coded as follows: 0 = no intervention needed; 1 = additional practice needed, 2 = additional instruction needed (and practice afterwards). The needs "additional instruction" and "additional instruction and practice" were combined into one, as based on students' work, we were not able to determine which of the two was most suited (see section 2.2.2.3 in Chapter 2 for an elaborate explanation).

To determine the accuracy of students' regulation judgments, we first coded students' actual need for intervention, based on a coding scheme we developed that is described in detail in section 2.1 in the Supplementary Materials. In short, we distinguished the same three categories as for students' regulation judgments (i.e., 0 = no intervention needed; 1 = additional practice needed, 2 = additional instruction needed [and practice afterwards]), based on the time students needed to complete the task and on whether they made computational or procedural errors.

Students' absolute regulation accuracy was computed by taking the absolute difference between students' regulation judgment and their actual need for intervention. It ranged from 0 to 2, with values closer to zero indicating more accurate regulation judgments.

### **3.2.3.4. Second-order Judgments**

Directly after students made the monitoring judgment before self-scoring, they made a second-order judgment (SOJ) about their monitoring accuracy (SOJ-m) by answering the question "How confident are you that you made a correct estimation during the previous question (question number ...)?". Directly after the regulation judgments before and after self-scoring, students made a SOJ about their regulation accuracy (SOJ-r) by answering the question "How confident are you that you made a correct choice during the previous question (question number ...)?". These SOJ-m and SOJ-r questions were answered on a 6-point Likert scale. In line with Fritzsche et al. (2018) this scale was labeled with smiley faces, see Figure 3.1.

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

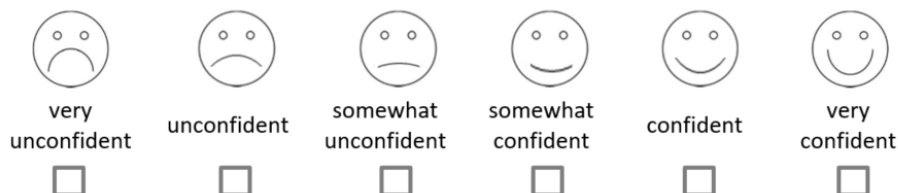
To check whether students understood the (second-order) judgment questions, we interviewed 12 of the pilot students (four low, four middle, and four high-performing students) one by one, after they completed the material. We asked them to describe the meaning of the (second-order) judgment questions on the multiplication task. All 12 students indicated that they understood the monitoring and regulation judgment questions. Eleven students correctly described the SOJ questions as their confidence in the 'correctness' of the previous judgment. One student described the SOJ questions as



their confidence in the 'correctness' of the previous judgment and in the performance on the multiplication task. Therefore, we decided that each time students had to answer a SOJ question, it was emphasized in the written question (see above) and by the experimenter to what previous question the SOJ question referred.

**Figure 3.1**

*Rating Scale of the Second-order Judgments.*



### 3.2.4. Procedure

After a short introduction by the experimenter, all students received the first booklet and a blue pen, and then started to complete the multiplication task. They were instructed to write down at what time they finished (the time was projected on the digital board in front of the class), but it was emphasized that there was no need to hurry (if students had mastered the content, 10 minutes should be enough, even without hurrying). When students finished the task, they were instructed to read the (fiction) books they kept in their drawers. After 12 minutes, the experimenter gave the instruction that the students who had not yet finished all problems should quit the task. Next, the students answered questions in their personal booklets (invested effort<sup>2</sup>, monitoring judgment, SOJ-m, regulation judgment, and SOJ-r). Each question was separately read aloud and explained by the experimenter. This procedure was then repeated for the division task. Next, all students received a second booklet and changed their blue pen for a green one. In the second booklet, students first self-scored their multiplication answers. Each problem was stated on a separate line together with the correct answer and with two boxes: "correct" and "incorrect or not answered." The experimenter explained that students had to look at their answers in the first booklet and tick the right box (the experimenter did not read the correct answers aloud). The following monitoring judgment (not used in the present study), regulation judgment and SOJ-r were again read aloud by the experimenter. This procedure of completing the second booklet was then repeated for the division task. This entire procedure (but with isomorphic problems) was repeated exactly one week later.

<sup>2</sup> The variable "invested effort" was not used in the present study but collected for use in other studies.

### 3.2.5. Analyses

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

Low-performing and high-performing students were defined separately for both tasks based on their average performance on the problems across the two days. In line with prior studies (Fritzsche et al., 2018; Händel & Dresel, 2018; Nederhand et al., 2021), low-performing students were defined as those scoring in approximately the first (lowest) quartile; high-performing students as those scoring in approximately the fourth (highest) quartile (note that these quartiles did not include the students who had answered all problems correctly or incorrectly; see section 3.2.2). On the six multiplication problems, low-performing students ( $n = 139$ ) correctly answered 0.5 to 2.5 problems on average. On the six division problems, low-performing students ( $n = 101$ ) correctly answered 0.5 to 1.5 problems. Thirty-nine students were defined as low performing on both the multiplication and division task. High-performing students answered 5.0 or 5.5 problems correctly on the multiplication task ( $n = 161$ ) and division task ( $n = 105$ ). Forty-one students were defined as high performing on both the multiplication and division task.

To answer RQ1 and 2 (on students' awareness of their monitoring and regulation [in]accuracy before self-scoring) we performed multilevel regression analyses in Mplus version 8 (Muthén & Muthén, 1998-2017), using maximum likelihood estimation with robust standard errors (MLR) which is robust to non-normality. We defined three levels in our data: day (level 1), student (level 2), and class (level 3). The class level was modeled using the "Complex" function, because we were not interested in the (fixed or random) effects on this level, we only wanted to account for the non-independence of observations within classes. To test RQ1 and 2, we analyzed how students' monitoring or regulation accuracy influenced their SOJs. As explained in section 3.1.2.1 in this Chapter and section 3.1 in the Supplementary Materials, we were mainly interested in the fixed effects at the day level (analyzing whether students, when they make more accurate judgments for the task on day 1 than for the task on day 2, feel more confident about their judgment accuracy for the task on day 1 compared to the task on day 2). To enable comparison with the results of Nederhand et al. (2021), the fixed effects at the student level are additionally reported in section 3.1 in the Supplementary Materials.

RQ3, regarding students' change in regulation accuracy awareness from before to after self-scoring, was explored based on the descriptive statistics, because the sizes of some sub-samples needed in these analyses were too small to perform statistical tests.

When data were missing because students did not complete a question (this applied to 0-6.7% per variable), data were deleted list-wise in the analyses. In each of the multilevel regression models, zero to four cases (a maximum of 2.1% of the data)

were identified as multivariate outliers. We were mainly interested in the results of the analyses without outliers to avoid drawing conclusions that are potentially affected by extreme cases in our data. For transparency we additionally ran the analyses also with outliers, which led to a difference in statistical significance for one of the fixed effects regarding RQ2b, for which we therefore report both effects, with and without outliers, in the Results section.

The data of this study are openly available in an online depository at [https://osf.io/36cak/?view\\_only=ba95fe7d0c6d4cd0a68a6bbbed76edf90](https://osf.io/36cak/?view_only=ba95fe7d0c6d4cd0a68a6bbbed76edf90).

### 3.3. Results

Descriptive statistics are presented in Table 3.1. The whole sample of students as well as the subsamples of low- and high-performing students made more accurate regulation judgments after self-scoring than before, on both tasks; however, for low- and high-performing students on the division task, this effect was small and not significant (for statistical analyses, see Oudman et al., 2022b). Before self-scoring, the whole sample of students and the subsets of low- and high-performing students felt, on average, 'somewhat confident' to 'confident' about their monitoring judgments (means between 4 and 5 out of 6) and felt 'confident' about their regulation judgments (means around 5). After self-scoring, students felt, on average, 'confident' to 'very confident' about their regulation judgments (means between 5 and 6; Table 3.1).

#### 3.3.1. Students' Monitoring Accuracy Awareness Before Self-scoring (RQ1)

Table 3.2 shows the results of the analyses in which students' SOJ-m/SOJ-r were regressed on their absolute monitoring/regulation accuracy prior to self-scoring, as an indication of students' monitoring/regulation accuracy awareness. For the whole sample of students, monitoring accuracy was not a significant predictor of students' SOJ-m on both tasks, indicating that on average, students were not aware of their monitoring (in)accuracy before self-scoring.

For the high-performing students, monitoring accuracy was not a significant predictor of students' SOJ-m on both tasks. For the low-performing students on the division task, monitoring accuracy was a significant and negative predictor of students' SOJ-m ( $B = -0.17$ ,  $p = .027$ ), but not on the multiplication task. On average, low-performing students' confidence in their monitoring (in)accuracy on the division task only increased with approximately 0.2 standard deviation when their monitoring accuracy increased with one standard deviation.

**Table 3.1***Means and Standard Deviations (Between Brackets) of Students' Performance, Monitoring/Regulation Accuracy, and Second-order Judgments*

Variable	Range	Multiplication				Division		
		Whole sample (N = 990)	LP students (n = 278)	HP Students (n = 322)	Whole sample (N = 718)	LP students (n = 202)	HP Students (n = 210)	
Performance	0 to 6	3.67 (1.84)	1.61 (1.38) <sup>b</sup>	5.26 (0.67) <sup>b</sup>	3.20 (2.06)	0.92 (0.83) <sup>c</sup>	5.30 (0.66)	
Before self-scoring								
Absolute monitoring accuracy <sup>a</sup>	0 to 6	1.23 (1.18)	1.67 (1.42) <sup>c</sup>	0.83 (0.80) <sup>c</sup>	0.99 (1.01)	1.06 (1.00)	0.77 (0.82)	
Absolute regulation accuracy <sup>a</sup>	0 to 2	0.61 (0.70)	0.63 (0.71) <sup>d</sup>	0.37 (0.57) <sup>d</sup>	0.43 (0.65)	0.33 (0.58)	0.25 (0.50)	
SOJ-m	1 to 6	4.51 (0.94)	4.38 (0.99) <sup>e</sup>	4.71 (0.93) <sup>e</sup>	4.56 (1.07)	4.30 (1.27) <sup>e</sup>	4.88 (0.89) <sup>e</sup>	
SOJ-r	1 to 6	4.91 (0.93)	4.95 (0.98)	5.02 (0.87)	5.06 (0.94)	5.06 (1.14)	5.17 (0.89)	
After self-scoring								
Absolute regulation accuracy <sup>a</sup>	0 to 2	0.44 (0.62)	0.47 (0.62) <sup>e</sup>	0.24 (0.47) <sup>e</sup>	0.36 (0.60)	0.28 (0.54)	0.20 (0.45)	
SOJ-r	1 to 6	5.29 (0.87)	5.22 (0.96) <sup>h</sup>	5.47 (0.68) <sup>h</sup>	5.30 (0.87)	5.34 (0.93)	5.45 (0.75)	

Note. LP = low-performing. HP = high-performing. Means are across both days. All means significantly differ from 0,  $p \leq 0.05$ .

<sup>a</sup>Values closer to zero indicate more accurate judgments. Other superscripts indicate significant differences between low- and high-performing students, tagged with the same letter,  $p \leq 0.05$ .

**Table 3.2**

*Effects of Absolute Monitoring/Regulation Accuracy on SOJ-m/SOJ-r Before Self-Scoring, at the Day Level*

	Whole sample		Low-performing students		High-performing students	
	<i>B (SE)</i>	<i>R</i> <sup>2</sup>	<i>B (SE)</i>	<i>R</i> <sup>2</sup>	<i>B (SE)</i>	<i>R</i> <sup>2</sup>
Monitoring						
Multiplication	0.00 (0.03)	.00	0.04 (0.04)	.00	-0.08 (0.07)	.01
Division	-0.07 (0.05)	.01	-0.17 (0.08)*	.02	-0.04 (0.06)	.00
Regulation						
Multiplication	-0.05 (0.05)	.00	-0.13 (0.10)	.01	-0.21 (0.08)**	.03
Division	-0.15 (0.05)***	.02	-0.36 (0.18)* <sup>a</sup>	.05	-0.03 (0.06)	.00

<sup>a</sup> Not significant when outliers were still included:  $B = -0.32 (0.18)$ ,  $p = 0.073$

\*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

### 3.3.2. Students' Regulation Accuracy Awareness Before Self-scoring (RQ2)

Across the whole sample, students' regulation accuracy was a significant and negative predictor of their SOJ-r, but only on the division task ( $B = -0.15$ ,  $p \leq .001$ ; Table 3.2). Students' confidence in their regulation accuracy on the division task only increased with approximately 0.1 standard deviation when their regulation accuracy increased with one standard deviation.

Regulation accuracy was also a significant and negative predictor of their SOJ-r for low-performing students on the division task ( $B = -0.36$ ,  $p = .039$ ) and for high-performing students on the multiplication task ( $B = -0.21$ ,  $p = .009$ ; Table 3.2). The increase in students' confidence in their regulation accuracy was approximately 0.2 standard deviation (for low-performing students on the division task and high-performing students on the multiplication task) when their regulation accuracy increased with one standard deviation.

### 3.3.3. Effect of Self-Scoring on Students' Regulation Accuracy Awareness (RQ3)

To explore the effect of self-scoring on students' regulation accuracy awareness, we looked at the patterns of change in students' SOJ-r from before to after self-scoring across four different subsets of students, of whom the regulation accuracy (1) increased after self-scoring, (2) decreased after self-scoring (3) was maximally accurate both before and after self-scoring, and (4) was equally inaccurate before and after self-scoring. Table 3.3 presents the change in students' SOJ-r from before to after self-scoring, for these four subsets of students. As shown in this table, for students in the subgroups whose accuracy increased or stayed maximally accurate, their confidence about their judgment accuracy increased,

which is desirable (as this might increase the likelihood that they act upon their regulation judgments). However, for most students in the subgroups whose accuracy decreased or stayed equally inaccurate, their confidence also increased (except for the high-performing students on the division task), which is not desirable.

Note though, that Table 3.3 also suggests that there are differences in the size of the increase in confidence: For the whole sample and for the subset of high-performers, on both tasks, the SOJ-r increase was on average larger for students whose regulation accuracy increased or stayed maximally accurate, compared to the students whose regulation accuracy decreased or stayed equally inaccurate. Thus, even though an increase in confidence was observed in all subgroups, this increase was smaller when it was undesirable. For the subset of high-performing students on the division task the pattern seems especially desirable, as within this subgroup, students' SOJ-r on average increased if students' regulation accuracy increased or stayed maximally accurate and *decreased* if their judgments became less accurate or stayed equally inaccurate after self-scoring.

**Table 3.3**

*Mean Change in Students' SOJ-r (and Standard Deviations Between Brackets) From Before to After Self-Scoring*

	<b>Accuracy increased after self-scoring</b>	<b>Accuracy decreased after self-scoring</b>	<b>Accuracy = 0 before and after self-scoring</b>	<b>Accuracy = 1 or 2 before and after self-scoring</b>
Multiplication				
Whole sample	<i>n</i> = 199 0.44 (1.34)	<i>n</i> = 74 0.36 (1.11)	<i>n</i> = 421 0.44 (0.88)	<i>n</i> = 242 0.21 (0.95)
LP students	<i>n</i> = 62 0.56 (1.47)	<i>n</i> = 27 0.30 (1.20)	<i>n</i> = 113 0.27 (0.96)	<i>n</i> = 67 0.12 (0.88)
HP students	<i>n</i> = 41 0.58 (1.16)	<i>n</i> = 9 0.33 (1.22)	<i>n</i> = 197 0.44 (0.72)	<i>n</i> = 58 0.34 (0.97)
Division				
Whole sample	<i>n</i> = 84 0.32 (1.19)	<i>n</i> = 45 0.09 (1.10)	<i>n</i> = 392 0.23 (0.85)	<i>n</i> = 129 0.22 (0.91)
LP students	<i>n</i> = 20 0.60 (1.10)	<i>n</i> = 10 0.50 (1.17)	<i>n</i> = 127 0.22 (0.77)	<i>n</i> = 30 0.37 (0.85)
HP students	<i>n</i> = 18 0.83 (1.29)	<i>n</i> = 14 -0.29 (1.13)	<i>n</i> = 138 0.30 (0.72)	<i>n</i> = 20 -0.05 (0.76)

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

Yet for the subset of low-performing students, we do not see such a clear pattern, although within the subset of low-performing students the SOJ-r increase is on average larger for students whose regulation accuracy increased compared to the students whose regulation accuracy decreased (for both the multiplication and division task). Likewise, low-performing students whose regulation judgments stayed maximally accurate showed a larger SOJ-r increase than low-performing students whose regulation judgments stayed equally inaccurate, but this was only true for the multiplication and not for the division task

## **3.4. Discussion**

Students' awareness of their monitoring and regulation (in)accuracy could help students in effective self-regulated learning behavior (e.g., seeking help when being unconfident of their accuracy). Nevertheless, monitoring accuracy awareness has hardly been investigated and regulation accuracy awareness not at all. In the present study, we explored to what extent primary school students (9-10 years old) showed awareness of their monitoring and regulation (in)accuracy on mathematical problem-solving tasks, how this differed between low- and high-performing students, and how students' regulation accuracy awareness was affected by self-scoring.

### **3.4.1. Are Students Aware of Their Monitoring and Regulation (In)accuracy Before Self-Scoring? (RQ1 and 2)**

First, we analyzed whether primary school students were able to distinguish between their more and less accurate monitoring judgments in terms of their confidence in those judgments, as an indication of monitoring accuracy awareness. Previous studies concluded that university students are somewhat aware of their monitoring (in)accuracy (Fritzsche et al., 2018; Händel & Dresel, 2018; Nederhand et al., 2021), but that secondary school students were not (Nederhand et al., 2021). This might suggest that monitoring accuracy awareness is a metacognitive skill that only develops during adulthood or late adolescence, in which case one would not expect to see it in primary school students. However, it could have been the case that accuracy awareness is related to task performance, and is found in students who perform high on a task but not in students who perform low on a task: Studies with university students (Fritzsche et al., 2018; Händel & Dresel, 2018) found that high-performing students were aware of their monitoring (in)accuracy and their low-performing peers were not. Our findings, however, seem to provide more evidence for the metacognitive development explanation, as the primary school students in our sample showed limited awareness of their monitoring (in)accuracy prior to self-scoring, across the whole sample (RQ1a) and also across the subsets of low- and high-performing students (RQ1b).

Note that our findings for RQ1 were based on analyses at the *within*-student level, therewith answering the question: when students make more accurate monitoring judgments for the task on day 1 than for the task on day 2, do they feel more confident about their judgment accuracy for the task on day 1 than for the task on day 2? Our hypothesis that primary school students would not be aware of their monitoring accuracy, was based on Nederhand et al. (2021) who found that secondary school students' monitoring accuracy did not predict their feeling of confidence in their accuracy at the *between*-student level. Interestingly, analyzing our data at the between-student level would not unambiguously lead to the conclusion that students' monitoring/regulation accuracy does not predict their feeling of confidence in their accuracy (see section 3.1 in the Supplementary Materials). But again, analyses at the between-student level answer a slightly different question (i.e., do students, who make a more accurate judgment on a task, feel more confident about the accuracy of this judgment, than students who make a less accurate judgment on that task?) than the one we were interested in. Future research should investigate whether the metacognitive development explanation holds, by testing effects at the within-student level across different age cohorts from the end of primary school until adolescence.

This was the first study to not only investigate students' awareness of their monitoring accuracy, but also of their regulation accuracy. Our primary school students showed also very limited awareness of their regulation (in)accuracy prior to self-scoring, across the whole sample (RQ2a) and across the subsets of low- and high- performing students (RQ2b). It is possible that regulation accuracy awareness also develops only at a later age, or it could be somehow dependent on monitoring accuracy awareness, which could be tested in future research with (young) adults.

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



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

### 3.4.2. What is the Effect of Self-Scoring on Students' Regulation Accuracy Awareness? (RQ3)

It is important that if students make accurate regulation judgments as result of an intervention, they also feel confident about their accuracy, because otherwise, students might not act upon their accurate judgments. In case students (still) make inaccurate judgments after self-scoring, it might be helpful if they feel relatively less confident about the accuracy of their judgments, as then they might adjust their initial judgments (with help of others).

On average, students' confidence in their regulation accuracy increased after self-scoring. However, this was not only the case for those students whose regulation accuracy increased or stayed accurate—which is desirable—, but also for students whose regulation accuracy decreased or stayed inaccurate—which is not desirable. This general increase in students' confidence of their regulation accuracy might perhaps be a consequence of the fact that students presumably felt highly confident about their monitoring accuracy after self-scoring (which we did not measure, but the pilot study strongly suggests that this is likely, see section 3.2.3.4). Self-scoring gives students information about the correctness of their answers, so this becomes a very salient cue to them and it is also highly diagnostic (as the students were quite accurate in self-scoring their work, resulting in very accurate monitoring judgments after self-scoring; Oudman et al., 2022b). Students' (implicit) knowledge that this cue (i.e., the self-rated correctness of their answers) was diagnostic could have resulted in feeling highly confident about their monitoring accuracy and in turn, to an increased feeling of confidence in their regulation accuracy after self-scoring—regardless of whether their regulation accuracy had *actually* increased. Nevertheless, students' confidence in their regulation accuracy after self-scoring increased more for students whose judgments indeed became more accurate or stayed accurate after self-scoring, than for students whose regulation accuracy became more inaccurate or stayed inaccurate. So, based on these findings, one could cautiously conclude that on average, self-scoring seemed to have a positive effect on students' awareness of their regulation accuracy, especially for students whose regulation judgments became more accurate or stayed accurate after self-scoring (RQ3a).

Finally, we addressed how self-scoring affected the differences between low- and high-performing students' awareness of their regulation (in)accuracy. Self-scoring seemed

to have affected the regulation accuracy awareness of the high-performing students more beneficially than of the low-performing students (i.e., especially for the high-performing students the confidence increase was larger for students who became more accurate or stayed maximally accurate than for students whose regulation accuracy became more/stayed inaccurate; RQ3b). One possible but tentative explanation for this difference might be that low-performing students were prone to wishful thinking, expressed by inflated confidence ratings. With regard to other metacognitive processes, such as predicting future performance, wishful thinking has also been found to be stronger for low- than for high-performing students (Serra & DeMarree, 2016). Future research should address whether this explanation holds.

Finally, future research should further investigate how students' regulation accuracy awareness, especially that of low-performing students, can be fostered, beyond self-scoring. Intervention studies could try to increase students' accuracy awareness by giving them feedback about their monitoring and regulation (in)accuracy, increasing students' knowledge about cue diagnosticity, and help student to give them insight into their own cue use.

### 3.4.3. Limitations

The current study was the first to investigate primary school students' awareness of their monitoring and regulation (in)accuracy. This study has several limitations. First, a potential limitation is that our measures of accuracy awareness differed somewhat from two prior studies on accuracy awareness amongst adults, which measured this construct by asking for item-specific judgments and analyzing to what extent students' SOJ-m were higher for accurate than for inaccurate monitoring judgments (e.g., Fritzsche et al., 2018; Händel & Dresel, 2018). In contrast, we asked students for whole-task judgments (and so did Nederhand et al., 2021), measuring students' accuracy on an interval scale, in our case ranging from zero to six (instead of accurate vs. inaccurate). Making judgments at this intermediate grain size, at which students judge the extent to which they master a specific skill, is regularly requested of students in primary education and can be useful when students reflect on which specific skills ask for an intervention (Hartwig & Dunlosky, 2017). Nevertheless, primary school (and older) students could also be asked to make item-specific (second-order) judgments and future studies could consider to investigate the effects of the grain size of (second-order) judgments on whether or not students are aware of their monitoring and regulation (in)accuracy.

Second, our findings only apply to procedural mathematics tasks, so future research could investigate to what extent these findings also apply to other subjects and tasks. Finally, our conclusions about the effect of self-scoring on students' regulation accuracy awareness have to be interpreted with caution, because they are based on exploring descriptive statistics. Future research should confirm these findings, ideally by conducting statistical tests, which would require larger subgroup sample sizes.

### **3.4.4. Implications and Conclusions**

When aiming to improve the effectiveness of students' self-regulated learning, it seems important to not only focus on improving their monitoring and regulation accuracy, but also on increasing their accuracy *awareness*. The present study indicates that (without intervention) primary school students have limited awareness of their monitoring and regulation (in)accuracy. Asking students to self-score their answers seemed to improve their regulation accuracy awareness overall, but this effect was limited for low-performing students and for students whose regulation accuracy decreased or stayed equally inaccurate after self-scoring. Future research on additional or other means to increase students' accuracy awareness (e.g., feedback or training) is needed, and ultimately, future research should address the question of whether the effectiveness of students' self-regulated learning indeed benefits more from intervention that not only focus on improving monitoring and regulation accuracy but also on improving students' accuracy awareness.





# Chapter 4

## Effects of Different Cue Types on the Accuracy of Primary School Teachers' Judgments of Students' Mathematical Understanding

This chapter was published as: Oudman, S., van de Pol, J., Bakker, A., Moerbeek, M., & van Gog, T. (2018). Effects of different cue types on the accuracy of primary school teachers' judgments of students' mathematical understanding. *Teaching and Teacher Education, 76*, 214-226. <https://doi.org/10.1016/j.tate.2018.02.007>

Author contributions: SO, JvdP, AB, and TvG designed the study. SO recruited participants, collected and analyzed the data, and drafted the manuscript. MM assisted in analyzing the data. SO, JvdP, AB, and TvG contributed to critical revision of the manuscript. JvdP, AB, and TvG supervised the study.

Acknowledgements: The authors would like to thank Susan Ravensbergen for her help with data collection and analyses.

## Abstract

To gain insight into how teachers' judgment accuracy can be improved, we investigated effects of cue-type availability. While thinking aloud, 21 teachers judged their students' ( $N = 176$ , 9-10 years old) decimal magnitude understanding. Sensitivity (correctly judging what students did understand) did not improve from availability of both answer cues<sup>1</sup> (students' answers to prior practice problems) and student cues (knowledge of students triggered by knowing their names), and was lower when only answer cues were available, compared to only student cues. Specificity (correctly judging what students did not understand) was higher when only answer cues were available, compared to only student cues or both student and answer cues.

---

1 In the other Chapters of this dissertation, answer cues are referred to as performance cues.

## 4.1. Introduction

To stimulate students' learning optimally, teachers need to provide adaptive instruction; that is, they have to tailor their explanations and instruction to a student's current level of understanding (e.g., Van de Pol et al., 2010). For teachers to be able to make adaptive instructional decisions, their judgments of their students' understanding need to be accurate (Klug et al., 2013; Südkamp et al., 2012; Van de Pol et al., 2011). Prior studies have shown, however, that there is much room for improving teachers' judgment accuracy (see for a meta-analysis Südkamp et al., 2012). This especially applies to teachers' judgment accuracy of students' conceptual mathematical understanding (Thiede et al., 2015). Yet, research that gives insight into how teachers' judgment accuracy can be improved is scarce.

Therefore, the first aim of the present study was to investigate how teachers' judgment accuracy of students' conceptual mathematical understanding can be enhanced, by manipulating the availability of information that can be used while making a judgment. According to the cue-utilization approach, judgments are based on specific pieces of information (i.e., cues) that can be more or less predictive (i.e., diagnostic) of students' actual understanding (Brunswik, 1956; Koriat, 1997; Thiede et al., 2010; Van Loon et al., 2014). The more predictive the cues being used, the more accurate a teacher's judgments of students' understanding will be. Manipulating which information is available will provide insight into which cues do and do not improve judgment accuracy. The second aim of the present study is to explore what cues teachers base their judgments on under the different cue-availability conditions, to gain more insight into their judgment process. This may ultimately aid the development of support tools to improve teachers' judgment accuracy.

### 4.1.1 Teachers' Judgments of Students' Conceptual Mathematical Understanding

In their meta-analysis Südkamp et al. (2012) conclude that teachers' judgment accuracy, reflected by the correlation between teachers' judgments of students' performance in language and mathematics and students' actual test performance, was positive and fairly high (Fisher's  $z$  transformed correlation = .63), but that there is still much room for improvement. As in most studies on teachers' judgment accuracy, in the studies included in the meta-analysis teachers' judgments were measured by asking teachers for one global rating per student (e.g., ratings of students' reading performance or a prediction of the number of correct answers on a test) or student rankings (e.g., ranking of the students in their class from lowest to highest mathematical understanding). The accuracy of these global judgments reflects teachers' knowledge on students' overall performance, not how well they are able to judge what individual students do and do not understand within a domain. Item-specific judgments do reflect this latter type of knowledge, which is what



teachers need in order to make adaptive instructional decisions, such as differentiating tasks or providing adequate instruction and feedback to individual students (Artelt & Rausch, 2014; Gabriele et al., 2016).

The few studies that did include item-specific judgments of students' mathematical understanding (Artelt & Rausch, 2014; Gabriele et al., 2016; Karing et al., 2011) found average "hit rates" (i.e., the proportion of accurately judged items when judging all items of a test) between the 58% and 78%. Taking into account that a random item prediction has on average 50% chance of being accurate, as teachers only indicated whether an item was answered correctly or incorrectly by the student, 58% is just above chance (i.e., when guessing). The item-specific judgments included in these studies did not distinguish between students' procedural skills and conceptual understanding in mathematics. Thiede et al. (2015) did make this distinction and their findings indicate that especially judging student conceptual mathematics understanding is challenging; they found that the average judgment accuracy for students' conceptual mathematics understanding, as measured by the gamma correlation (computed across the students within a class), was only .20 after intervention (vs. gamma correlation = .66 for computational skills; 1 would mean perfect prediction). In sum, prior studies showed there is a need to improve the accuracy of teachers' (item-specific) judgments of students' conceptual mathematics understanding, but knowledge on how to do so is lacking.

#### **4.1.1.1. Teachers' Cue Utilization**

Teachers make numerous instructional decisions during everyday teaching practice that are based on judgments of their students' current level of understanding. The accuracy of these judgments could be influenced by several factors, such as teacher characteristics (e.g., their professional expertise) and characteristics of the test that is used to make judgments (e.g., the subject area; Südkamp et al., 2012). In the current study we especially focus on the specific pieces of information that teachers base their judgments on, typically referred to as cues (Brunswik, 1956; Koriat, 1997). For instance, a teacher's observation that a particular student flawlessly completed yesterday's assignment on decimal magnitude (i.e., a cue) can lead to the judgment that this student's understanding of decimal magnitude is excellent. In turn, the teacher may make the instructional decision that the student can skip today's exercises on decimal magnitude and continue with exercises on adding decimals.

Studies on what cues teachers actually use when judging students' understanding are scarce and differ strongly in their methodology. Whitmer (1983) interviewed elementary teachers about the information they commonly used when giving grades for mathematics and language. Webb (2015) also interviewed elementary teachers, directly after they made prospective predictions of how their students scored on a mathematics test, and asked them what they based their predicted scores on. Cooksey et al. (2007) asked elementary

teachers to think aloud while making retrospective judgments of students' written texts (i.e., the teachers were provided with students' products). Cues that were frequently reported by the teachers in these three studies were students' prior performance in a specific subject; students' general cognitive abilities and learning disorders; students' problem-solving skills; students' motivation and interest; students' effort and discipline; what content had been taught or practiced previously; and the difficulty of a specific domain or task. Apparently, teachers derive the cues they use to inform their judgments from different information sources, such as the content material or characteristics of the task at hand (i.e., task content cues), information about students' prior performance (i.e., answer cues), and more general information about the students (i.e., student cues).

Several studies have zoomed in on teachers' use of student cues, rather than on other cue types. Correlational studies showed that students' ethnicity, SES, classroom engagement, disability status, and social competency were predictive of the height of teachers' judgments of their students' literacy and mathematical understanding, even when controlling for students' actual performance (e.g., Furnari et al., 2017; Hurwitz et al., 2007; Kaiser et al., 2013; Paleczek et al., 2017; Ready & Wright, 2011). This implies that teachers use these student cues while making judgments of students' understanding. Comparable conclusions can be drawn from two experimental studies by Kaiser and colleagues (2013, 2017), in which participants directed pre-designed questions at fictional students and observed their responses in a simulated classroom environment. Next, participants indicated how many of those pre-designed questions they thought each of the students had answered correctly, reflecting teachers' judgments of students' mathematics or reading achievement. Students' engagement (operationalized as the probability of a simulated student volunteering to answer a question; Kaiser et al., 2013) and students' minority status (Kaiser et al., 2017) were significantly related to teachers' judgments of the amount of correctly answered questions. In another study, Kaiser et al. (2015) compared the effects of different types of information on pre-service teachers' judgment accuracy of fictional students' mathematics grades. All teachers were provided with information on students' oral and written achievement in mathematics and some teachers were additionally provided with student characteristics such as students' self-concept and intelligence. From the significant correlation between teachers' judgments of the mathematics grade and the value of the presented student characteristics (i.e., gender, intelligence, and German dictation exercise grade) it can again be concluded that teachers probably used these student cues while making judgments of students' mathematics performance.

#### ***4.1.1.2. Relation Between Teachers' Cue Utilization and Judgment Accuracy***

As mentioned before, the more predictive the cues that the teacher uses in making judgments of a student's understanding, the higher the judgment accuracy will be

(Brunswik, 1956; Koriat, 1997; Thiede et al., 2010; Van Loon et al., 2014). For example, the correctness of a student's conceptual expression during a class discussion might be more predictive for this students' actual understanding of the content material than the time the student spends on a task. Basing judgments on more predictive cues will lead to more accurate judgments, which in turn will lead to more adaptive instructional decisions (Klug et al., 2013; Südkamp et al., 2012; Van de Pol et al., 2011).

In the experimental study by Kaiser et al. (2015) teachers who were only provided with information on students' oral and written achievement in mathematics, made more accurate judgments of fictional students' mathematics grades than teachers who were additionally provided with student characteristics (i.e., students' engagement, minority status, gender, intelligence, and German dictation exercise grade). This finding suggests that student cues might be not predictive of students' actual understanding, as the availability (and therefore presumably, the use) of such cues resulted in less accurate judgments of students' understanding. It is an open question whether this would also apply when teachers make judgments of their own students (of whom they might—in addition to such non-predictive cues—also have knowledge that could be predictive).

Moreover, the question on what cues teachers should focus to increase the accuracy of their judgments has hardly been addressed to date. Research on student judgments of their own understanding has shown that redirecting students' attention to products of generative activities (see Fiorella & Mayer, 2015) improved students' judgment accuracy of their text understanding compared to students who were not encouraged to engage in such generative activities. The generating activities consisted of generating keywords (De Bruin et al., 2011), self-explanations (Griffin et al., 2008), making diagrams (Van Loon et al., 2014), writing summaries (Thiede et al., 2010), or making concept maps (Thiede et al., 2010).

Teachers can also obtain cues from students' answers that result from written or oral generative activities (i.e., answer cues). Some evidence that focusing on answer cues improves teachers' judgment accuracy of students' mathematical understanding comes from a study by Thiede et al. (2015). They examined whether teachers' judgment accuracy of students' mathematical understanding was affected by involvement in a professional development program. This program stimulated teachers to focus more on student products that give insight into student thinking (e.g., by asking students to articulate their way of reasoning) during teaching. Teachers who took part in the program indeed made more accurate judgments of students' mathematical computational skills and conceptual understanding than teachers who did not participate in the training program. Nevertheless, judgment accuracy for students' conceptual understanding was still quite poor after participation (gamma correlation: .20; 1 would mean perfect prediction). Besides, it remains unclear whether it was indeed the increased focus on answer cues, or some other aspects of the 45-hour training that caused the improvement in teachers' judgment accuracy (e.g., improved mathematical content knowledge).

### 4.1.2 Present Study

Accurately judging students' conceptual mathematics understanding seems a challenging task that needs further investigation. In the present study, we experimentally investigated whether giving teachers access to cues with a high expected predictive value (i.e., answer cues)—compared to student cues—would improve teachers' judgment accuracy of students' conceptual understanding of decimal magnitude. More specifically, the first Research Question is whether teachers' judgment accuracy is affected when answer cues are available, additional to or instead of student cues, compared to when only student cues are available.

To answer this question, we experimentally manipulated the availability of the different cue types by providing teachers with a name of a student from their own class (student cues only), the anonymized answers on decimal magnitude practice problems of one of their students (answer cues only), or both the student's name and his/her answers (student + answer cues). We measured teachers' judgment accuracy of students' decimal magnitude understanding by comparing teachers' item-specific predictions of how well their students would perform on a decimal magnitude test with students' actual performance on such a test. The material (i.e., the first assignment consisted of practice problems and the second assignment consisted of items of which teachers had to predict students' performance) was created in such a way that analysis of students' answers on the first assignment could provide teachers with information on students' (mis)conceptions in the domain of decimal magnitude. An example of such a misconception is that students think of decimals as if they are whole numbers (e.g., 0.35 is greater than 0.8; see Durkin & Rittle-Johnson, 2015; Isotani et al., 2011).

With regard to the first Research Question we hypothesized that teachers' judgment accuracy would be lower in the name-only (i.e., student cues) condition than in both the answers-only (i.e., answer cues) and the name+answers (i.e., student + answer cues) condition, because answer cues can be expected to be more predictive of students' performance on the test assignment than the more general student characteristics on which teachers had to rely in the name-only condition (see section 4.1.1.2). When judging their own students instead of fictional students (the latter was the case in the study by Kaiser et al., 2015), teachers may have knowledge about their students that could be predictive of students' actual understanding (e.g., knowledge on students' general conceptual mathematics understanding). However, knowing the student's name will likely also activate non-predictive student cues, whereas students' answers on practice problems are more directly associated with their understanding and as a result more predictive.

Given that simulated classroom research with fictional students showed that teachers' judgment accuracy was impaired when student characteristics were available (Kaiser et al., 2015), the second question we addressed is whether the accuracy of teachers' judgments of their own students' understanding is affected when only answer cues are available compared to when both student and answer cues are available. Focusing on

student cues with presumably low predictive value might interfere with thorough or full analysis of students' answers on the first assignment. Hence, the second hypothesis we test in the present study is that teachers make more accurate judgments in the answers-only condition than in the name+answers condition.

As we expect that hypothesized differences in judgment accuracy between conditions would be due to differences in teachers' cue use across conditions (see section 4.1.1.2), the second aim of this study is to explore differences in teachers' cue utilization between conditions. The third Research Question we addressed is: (How) do the cues that teachers use when making judgments of students' decimal magnitude understanding differ across the name-only, name+answers and answers-only conditions?

## 4.2. Methods

### 4.2.1. Participants

#### 4.2.1.1. Teachers

Twenty-one teachers (17 female) from 17 different primary schools in the Netherlands, teaching 9- to 10-year-old students (i.e., Dutch grade six, comparable to US grade four in terms of age) volunteered to participate in this study. They were between 25 and 54 years old ( $M_{\text{age}} = 36.34$ ,  $SD = 8.99$ ) and had between three and 33 years of teaching experience ( $M = 10.33$ ,  $SD = 7.60$ ). They had been teaching their classes between two and five days a week ( $M = 3.88$ ,  $SD = 1.24$ ) from the beginning of the school year (i.e., end of August; data collection took place in October and November 2016). Six of them had been teaching the students in their class in a previous grade as well. Eight teachers had completed additional mathematics education courses after graduating from regular teacher training (e.g., on serious mathematics problems/dyscalculia or courses required to become the school's mathematics specialist).

#### 4.2.1.2. Students

Out of the 454 students who attended the 21 participating classes, 418 were included in the study (224 girls,  $M_{\text{age}} = 9.55$ ,  $SD = 0.42$ ). Students were excluded because of following special mathematics programs ( $n = 20$ ), no parental consent to use students' data ( $n = 11$ ), large portions of incomplete assignments ( $n = 2$ ), or because their teachers accidentally saw their answers during task completion ( $n = 3$ ). From this sample, three students in each condition (hereafter: "target students") were selected per teacher (i.e., nine in total per teacher). Based on students' test performance, a low, medium, and high performing student was selected per condition (see section 4.2.5). Due to time restrictions (see section 4.2.5) 13 of the target students (max. 2 per teacher) were dropped from the procedure. This resulted in a final sample of 176 students about whom the teachers made judgments (82 girls;  $M_{\text{age}} = 9.59$ ,  $SD = 0.42$ ; name-only condition:

$n = 62$ ; name+answers condition:  $n = 57$ ; answers-only condition:  $n = 57$ ). At the time the study took place, decimal magnitude had not yet been taught (this is not done before the end of Dutch sixth grade), so the topic was new to almost all students.

### 4.2.2. Design

This study had a within-subjects design, with all 21 teachers making judgments of students' decimal magnitude understanding under three conditions: 1) name only (teachers were only provided with student names), 2) name+answers (teachers were provided with student names and students' answers to prior practice problems), and 3) answers only (teachers were provided with anonymized answers only). Teachers made judgments of three students per condition while thinking aloud.

### 4.2.3. Materials

Students were provided with instructions and assignments in the domain of decimal magnitude. Student (mis)conceptions in the domain of decimal magnitude are clearly defined. Five common and persistent misconceptions are: (1) thinking of decimals as if they are whole numbers (e.g., 0.35 is greater than 0.8 because 35 is greater than 8); (2) ignoring a zero that is in the tenths place (e.g., 0.08 is the same as 0.8); (3) assuming that adding a zero at the end of the decimal increases its magnitude (e.g., 0.30 is greater than 0.3); (4) viewing decimals less than one as being less than zero or more than one (e.g., 0.2 is less than 0); and (5) treating decimals as fractions thus thinking that numbers with more decimals are smaller (e.g., 0.852 is smaller than 0.3; see Durkin & Rittle-Johnson, 2015; Isotani et al., 2011).

#### 4.2.3.1. Introductory Video Lesson

In an introductory video lesson, the topic of decimal magnitude was introduced to the students by explaining the place values of the tenths, hundredths and thousandths on a number line by connecting its meaning to fractions. No explicit attention was paid to specific misconceptions. Moreover, the study procedure was explained to the students in the video. The video had a total duration of 8:30 minutes. This video was created by the first author—who is also a primary school teacher—based on the most commonly used Dutch mathematics textbooks.

#### 4.2.3.2. Student Assignments

Students' answers on the first assignment (i.e., practice problems) functioned as a product of student generative activities that may give insight into student thinking. The first assignment consisted of 16 number line problems on decimal magnitude (nine multiple choice and seven open problems). The assignment was constructed such that each wrong answer was indicative of a particular misconception. For instance, when students placed 0.07 near the location of 0.7 on the number line, they were considered to hold the

“ignoring the zero in the tenths place” misconception (the complete assignment, including indication of the misconceptions, is provided as online supplementary material; section 4.1 of the Supplementary Materials). For some items, multiple answer options indicated the same misconceptions. For other items, different answer options indicated a different misconception. In total, each misconception could become evident four or five times. The items were based on examples from earlier research about student misconceptions in the field of decimal magnitude (Adams et al., 2014; Durkin & Rittle-Johnson, 2012, 2015; Rittle-Johnson et al., 2001).

The second assignment (i.e., test problems) consisted of 17 decimal magnitude problems; five number line problems and 12 word problems (15 multiple choice and two open problems; all included as online supplementary material; section 4.1 of the Supplementary Materials). Because the format of the items differed substantially between the first and second assignment, teachers had to use their interpretations of student thinking (i.e., students’ misconceptions) when making judgments in the name+answers and answers-only condition (i.e., they could not directly translate correctness of an item in the first assignment into a judgment on the correctness of a particular item in the second assignment). The items in the second assignment were also based on examples from earlier research about student misconceptions in the field of decimal magnitude (Adams et al., 2014; Durkin & Rittle-Johnson, 2012, 2015; Rittle-Johnson et al., 2001). In total, each misconception could become evident three to five times. Two of the items were considered to assess students’ overall understanding of decimal place values. Students’ performance on the second assignment was scored by assigning one point for each correct answer (min = 0, max = 17).

Correlations between each of the misconceptions at the first assignment and the same misconception at the second assignment (measured by the number of errors indicative of the misconception) were significant and ranged from low to moderate ( $r_{\text{whole number}} = .51$ ,  $r_{\text{ignoring zero in tenths place}} = .49$ ,  $r_{\text{fraction}} = .41$ ,  $r_{\text{outside 0 and 1}} = .27$ ,  $r_{\text{zero at end makes bigger}} = .52$ , for all  $p < .001$ ), meaning that the answer cues on the first assignment have (modest) predictive value for performance on the second assignment.

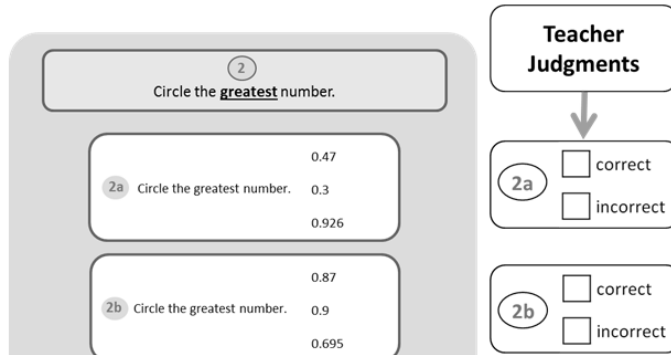
#### 4.2.4. Teachers’ Judgments

Teachers were asked to make item-specific judgments about the performance on the second assignment of the nine target students from their classroom; they saw the 17 test items and indicated for each item whether they thought that the student had answered it correctly or incorrectly (see Figure 4.1). For the three students in the name-only condition, teachers had to make these item-specific judgments knowing only the name of the students. For the three students in the name+answers condition, teachers were provided with students’ names and students’ answers on the practice assignment. Finally, for the three target students in the answers-only condition, teachers had to make

the judgments seeing only students' answers on the practice assignment. Note that the completed practice assignment did not trace back to specific students, since students all used the same pencil type and were instructed not to write on the test sheets except for marking the place on the number line or answer option they thought was correct.

**Figure 4.1**

*Fragment of Teachers' Judgment Task (Translated From Dutch)*



Most prior studies treat teachers' judgment accuracy as a single process through which teachers both judge what students do understand and do not understand (e.g., Kaiser et al., 2015; Südkamp et al., 2012; Thiede et al., 2015). In line with recent studies on students' own judgment accuracy we applied a two-process model, focusing separately on judgments of what students do understand (called "sensitivity" or "certainty") and judgments of what students do not understand (called "specificity" or "uncertainty"; cf. Rutherford, 2017; Schraw et al., 2013). Knowing what students understand seems necessary for teachers to be able to anchor instructions and tasks to concepts and procedures already mastered by students; knowing what students do not understand (i.e., which misconceptions they have or where gaps in their knowledge lie) seems, for instance, necessary to give adequate additional instruction. Instruction will only foster conceptual change when it addresses the specific (mis)conceptions held by students (Prediger, 2008). Modeling both sensitivity and specificity allows examination of the potentially different processes surrounding accurate teacher judgments.

Sensitivity was calculated by first counting the number of items that were answered correctly by a student and were judged accurately (i.e., judged as correct) by the teacher. This number was then divided by the total number of items answered correctly by that student. For instance, when a student answered 10 of the 17 items correctly and the teacher judged five of the 10 correctly answered items accurately as being correctly answered, the sensitivity value was 0.5. Specificity was calculated by dividing the number of items that were answered incorrectly and were judged accurately by



the teacher as being incorrectly answered, by the total number of items answered incorrectly. Consequently, teachers received a score between 0 and 1 for sensitivity and specificity for each student, with 0 indicating that none of the correct (sensitivity) or incorrect (specificity) items was judged accurately, and 1 indicating that all of the correct (sensitivity) or incorrect (specificity) items were judged accurately. For 11 of the 176 students (6%) we could not compute a specificity measure because they answered all items correctly. Because those data were Missing Not At Random (MNAR), listwise deletion or methods such as multiple imputation could not be applied (Van Buuren, 2012). We decided to assign these students value 1 for specificity (the maximum value), since these students had zero incorrect answers and teachers as a matter of fact judged 100% of this number of incorrect answers also as incorrect. To check whether this decision affected the results, we additionally conducted the main analyses with listwise deletion (another method to deal with missing data) of the students who answered all items correctly. Listwise deletion led to the same pattern of results as those presented in section 4.3.1.1.

#### **4.2.5. Procedure**

The study procedure consisted of a student and a teacher part. The student part took place during a normal lesson day and lasted about 45 minutes. Students were informed that they would see an introductory video and make some tasks on the novel topic of decimal magnitude. They were also informed that they would not receive a grade for their work, but were encouraged to try their best on the assignments. Then, the introductory video was shown, after which students individually worked on the first and second assignment. Their teachers were present during the lesson, but had been instructed not to help students, answer questions, or look at students' answers (to prevent them from obtaining specific knowledge of some students' decimal magnitude understanding). Researchers only answered student questions that were not related to the mathematical content. After both assignments had been completed, the student work was collected and the second assignment of each student was scored by the researchers. To ensure that teachers judged students with varying understanding of decimal magnitude, we divided all students within one class into three groups. The groups were based on the expected amount of decimal magnitude misconceptions students held, as represented by their performance on the second assignment. We applied the following distinction: high score = 14 – 17 (expected to hold no or one misconception), medium score = 10 – 13 (expected to hold two misconceptions), and low score = 0 – 9 (expected to hold more than two misconceptions). Nine target students per teacher were then selected: three low, three medium, and three high scoring students. In each judgment condition, teachers would encounter one student with a high, one with a medium, and one with a low score. When there were not enough students in a score category, a student from

another category with the nearest score was selected (e.g., when there were not enough students in the high category a student from the medium category with a score of 13 was placed in the high category). The average test scores of the selected students were comparable across conditions;  $M_{\text{name-only}} = 10.68$  ( $SD = 3.61$ ),  $M_{\text{name+answers}} = 10.79$  ( $SD = 3.72$ ), and  $M_{\text{answers-only}} = 10.79$  ( $SD = 3.80$ ).

After the students went home, the teacher part started. Teachers first completed the first and second assignment themselves to become familiar with the assignments. Then, they judged the target students' performance on the second assignment, by condition: first for the three students in the name-only condition, then for the three students in the name+answers condition, then for the three students in the answers-only condition. Although the order of low/medium/high performing students was randomized within the conditions, the order of conditions was fixed to avoid that teachers would be triggered to use other cues in the name-only and name+answers conditions (e.g., related to observed student performance on other mathematics tasks) than they would normally do. Note that data from a prior study on text comprehension judgments (Van de Pol et al., 2017), showed no signs of a learning effect (i.e., teachers' judgment accuracy does not increase as they gain more experience with making judgments). Prior to the name-only and name+answers condition, the teachers were familiarized with the condition-specific judgment procedure in a practice phase with five test items that they had to judge for one of their non-target students. Because a pilot study had shown that teachers became less alert after one hour of judging, the judgment phase was ended after one hour. When the researchers, in the beginning of the judgment process, noticed that a participant was relatively slow, they decided to drop a student from the name+answers and/or answers-only condition (see for numbers section 4.2.1.2 or Table 4.2).

Teachers were asked to think aloud while making the judgments, to gain insight into teachers' cue utilization (cf. Cooksey et al., 2007). The participants were prompted to continue thinking aloud when they were silent for five seconds or more, but were not asked for clarifications or elaborations as this might interfere with the cognitive processes involved in making the judgments (Ericsson & Simon, 1993; Van Someren et al., 1994). Research has shown that thinking aloud does not affect marking processes (which are presumably closely related to judgment processes; Crisp, 2008) or change the course or structure of thought processes in general (Ericsson & Simon, 1993; Van Someren et al., 1994). Although think-aloud protocols can slow down the process and probably do not reflect *all* of a person's thoughts, they do provide more information on cognitive processes than most other methods such as prospective interviews or self-reports (e.g., Ericsson & Simon, 1993; Van Someren et al., 1994). The 21 think-aloud protocols were audio-recorded, transcribed, and anonymized.

## 4.2.6. Data Analysis

### 4.2.6.1. Analyses of Accuracy Differences Across Conditions

To investigate the effects of availability of only student cues, only answer cues, or both student and answer cues on teachers' judgment accuracy (Research Question 1 and 2) we performed a multilevel regression analysis in Mplus version 8 (Muthén & Muthén, 1998-2017). To account for the nested data structure with students (level 1) clustered in classes and thus in teachers (level 2), the "Complex" function in Mplus was used with maximum likelihood estimation with robust standard errors (MLR). A regression analysis with two outcome measures (i.e., sensitivity and specificity) was applied, because sensitivity and specificity correlated significantly with each other,  $r_{\text{zero-order}} = -.353, p < .001$ . The predictor variable condition was added using dummy coding.

### 4.2.6.2. Coding the Think-aloud Protocols

In order to investigate which cues teachers used (Research Question 3) we analyzed their think-aloud data. The 176 think-aloud transcripts were coded to identify the cues teachers reported while making their judgments.

First, to ensure a systematic segmentation procedure independent of coding categories, we defined a unit of analysis as "a sentence or part of a compound sentence that can be regarded as meaningful in itself, regardless of the meaning of the coding categories" (Strijbos et al., 2006, p. 37). A subsample of six transcripts (two from each condition, randomly selected) were independently segmented by two coders (the first author and a research assistant). The proportion agreement was determined from the perspective of each coder serving as an upper and lower bound of the 'true' agreement (cf. Strijbos et al., 2006). The proportion agreement had a lower bound of 88.9% and an upper bound of 90.4%, both above the threshold of 80% (cf. Strijbos et al., 2006). The coders respectively segmented the transcripts into 422 and 415 segments. In case of disagreement, the coders reached consensus on the segmentation through discussion.

The transcripts were coded in three steps. The final coding scheme including descriptions and examples can be found in section 4.2 of the Supplementary Materials. For the first step of open coding 12 transcripts (4 from each condition) were used, across which 64 different codes were identified. Next, we divided these codes into categories, resulting in a coding scheme of 26 categories. We then checked whether these categories sufficed by applying this coding scheme to 6 further transcripts (2 from each condition). To check the interrater reliability, the two coders independently coded 10% of the transcripts (18 transcripts, 6 from each condition, in total 1124 segments). The interrater reliability was sufficient to high ( $\kappa = .79$ , agreement = 81.9%; Landis & Koch, 1977). In case of disagreement, the coders reached consensus on the coding through discussion.

For the analyses, these 26 categories were aggregated into six main categories: Content, Student, Answers, Student\*Content, Teacher, and Miscellaneous. The first four

categories refer to the information sources teachers presumably used in their judgments; these four were included in the analyses. The Content category included codes of statements related to curriculum content and material content (e.g., statements about what was or was not yet taught in the curriculum thus far, or about item characteristics of the first or second assignment). The Student category included codes assigned to statements about student characteristics (e.g., statements related to students' general cognitive ability or students' motivation). One particular interesting subcategory of the Student category was "fabricated student", assigned to statements occurring in the answers-only condition implying that teachers had an idea about the identity of the student or tried to guess the identity. Codes within the Answers category were based on students' answers on the practice assignment (e.g., statements related to a student's performance on one item or a group of items). Codes within the Student\*Content category could be based on the student, the content material or the answers, or a combination of these, but always reflected an interaction between the student and the content (e.g., statements referring to a decimal misconception held or a strategy used by a student).

The two remaining categories (i.e., Teacher and Miscellaneous) included codes that were irrelevant for answering our research questions. The Teacher category included statements about teachers' emotions or meta-thoughts about the judgment process. The Miscellaneous category included all other irrelevant codes (e.g., unclear statements).

When teachers received the same code on multiple sequential segments, for example: "I think it took student x quite a while. I saw that it took her a long time", we would count this code only once. Hence, each segment was additionally coded with regard to repetitions. When one of the 26 codes from the coding scheme was repeated within one completed argumentation of a teacher (i.e., describing a student before starting with the judgments, or when analyzing a student's answer on one item or judging one item) this was also coded as "repetition". Repetitions were excluded from the frequency statistics as presented in the Results section and also excluded from the analyses. The reliability of applying the repetition codes was determined by independently coding the repetition dimension of 6 transcripts (2 from each condition, in total 511 segments) that were already segmented and coded with codes from the coding scheme by one of the coders. The interrater reliability for the repetition dimension was very high ( $\kappa = .93$ , agreement = 97.5%).

After reliability was checked, the rest of the data (including the data that was used for developing the coding scheme) was segmented and coded definitively. The two coders each coded half of the data. In cases of doubt about segmenting or what code to apply the coders reached consensus on the coding through discussion. After the coders each coded the transcripts of three teachers, they calibrated by independently coding three segmented transcripts and discussing the cases of disagreement until consensus, before continuing with the next three teachers.

### 4.2.6.3. Analyses of Differences in Cue Utilization Across Conditions

To investigate the effects of the availability of only student cues, only answer cues, or both student and answer cues on teachers' cue utilization (Research Question 3), we performed multilevel regression analysis comparable to one conducted to investigate the accuracy differences (see section 4.2.6.1). Instead of sensitivity and specificity the average frequencies of the four relevant main categories per student (excluding repetitions) were included as outcome measures.

## 4.3. Results

### 4.3.1. Teachers' Sensitivity and Specificity

Table 4.1 displays a cross tabulation of teachers' item specific judgments and students' actual item performance, including student and teacher totals. Table 4.2 displays teachers' average sensitivity and specificity values per condition. Teachers on average judged 8.15 (75%), 7.63 (69%), and 7.09 (64%) of students' correctly answered items as correct (i.e., sensitivity), and 2.63 (41%), 3.37 (48%), and 4.16 (64%) of students' incorrectly answered items as incorrect (i.e., specificity) in the name-only, name+answers, and answers-only condition, respectively.

**Table 4.1**

*Cross Tabulation of Teachers' Item-Specific Judgments and Students' Actual Test Assignment Scores, Including Student and Teacher Totals*

	<b>Student correct (SD)</b>	<b>Student incorrect (SD)</b>	<b>Total Teacher (SD)</b>
Name-only			
Teacher correct	8.15 (3.88)	3.69 (2.53)	11.84 (3.66)
Teacher incorrect	2.53 (2.28)	2.63 (2.75)	5.16 (3.66)
Total Student	10.68 (3.61)	6.32 (3.61)	10.77 <sup>a</sup> (2.78)
Name+Answers			
Teacher correct	7.63 (3.92)	2.84 (1.74)	10.47 (3.68)
Teacher incorrect	3.16 (2.23)	3.37 (3.14)	6.53 (3.68)
Total Student	10.79 (3.72)	6.21 (3.72)	11.00 <sup>a</sup> (2.41)
Answers-only			
Teacher correct	7.09 (4.11)	2.05 (1.69)	9.14 (4.00)
Teacher incorrect	3.70 (2.84)	4.16 (3.30)	7.86 (4.00)
Total Student	10.79 (3.80)	6.21 (3.80)	11.25 <sup>a</sup> (2.87)

*Note.* Numbers represent the absolute number of items judged as, or answered correctly/incorrectly.

<sup>a</sup> Average number of items (answered correctly and incorrectly by students) that teachers judged accurately.

**Table 4.2***Mean Sensitivity and Specificity Values per Condition*

Condition	N	Sensitivity (SD) <sup>a</sup>	Specificity (SD) <sup>a</sup>
Name-only	62	.75 (.20)	.41 (.31)
Name+answers	57	.69 (.22)	.48 (.32)
Answers-only	57	.64 (.24)	.64 (.31)

<sup>a</sup> min. = 0, max = 1**Table 4.3***Parameter Estimates from a Multilevel Analysis on Teachers' Sensitivity and Specificity*

Effects	B	SE	Cohen's d	p
Sensitivity				
Name-only vs. Name+Answers	0.06	0.03	0.27	.066
Name+answers vs. Answers-only	0.05	0.04	0.20	.265
Name-only vs. Answers-only	0.11	0.04	0.48	.014*
Specificity				
Name-only vs. Name+Answers	-0.08	0.05	-0.24	.120
Name+answers vs. Answers-only	-0.16	0.04	-0.49	< .001*
Name-only vs. Answers-only	-0.24	0.07	-0.73	< .001*

\* This effect significantly differed from zero when applying Bonferroni correction for multiple hypotheses testing, using an alpha level of  $0.05/3 = 0.017$ 

#### 4.3.1.1. The Effect of Cue-type Availability on Teachers' Sensitivity and Specificity

Table 4.3 displays the results of a multilevel analysis on sensitivity and specificity including condition as predictor. Regarding our first Research Question, we tested whether teachers' sensitivity and specificity was higher in the name+answers and answers condition, than in the name-only condition. Comparison of the name-only and the name+answers condition did not show significant differences in teachers' sensitivity and specificity in those conditions (sensitivity:  $p = .066$ ; specificity:  $p = .120$ ). Thus, gaining access to students' answers on practice problems did not significantly improve teachers' judgments of what students did and did not understand compared to when teachers could solely rely on their general knowledge of the students. Comparison of the name-only and the answers-only condition showed a significant difference in sensitivity ( $p = .014$ ), but not in the expected direction: sensitivity was higher in the name-only condition. The regression coefficient shows that teachers' sensitivity increased with 0.11

when teachers made judgments in the name-only condition compared to the answers-only condition and that the effect size (0.48) was small to medium (cf. Cohen, 1992). In line with our hypothesis, though, teachers' specificity was higher in the answers-only than in the name-only condition ( $p < .001$ ). Thus, teachers were more accurate at indicating what students did understand, but less accurate at indicating what students did not understand when they could only rely on general knowledge of their students (triggered by access to students' names) than when they could only rely on students' anonymized answers on practice problems. The regression coefficient shows that teachers' specificity increased with 0.24 when teachers made judgments in the answers-only condition, compared to the name-only condition and that the effect size (0.73) was medium to large.

Regarding the second Research Question, contrary to our hypothesis, the analysis showed that teachers' sensitivity in the answers-only condition did not differ significantly from the name+answers condition ( $p = .265$ ). In line with our hypothesis, however, teachers' specificity in the answers-only condition was significantly higher than in the name+answers condition ( $p < .001$ ). Thus, the teachers were better able to indicate what students did not understand, when they could only see students' answers on practice problems (i.e., anonymized) than when they knew the name of the student who produced these answers. The regression coefficient shows that teachers' specificity increased with 0.16 when teachers made judgments in the answers-only condition, compared to the name+answers condition and that the effect size (0.49) was small to medium.

### 4.3.2. Cues Reported by Teachers

In Table 4.4, all cues reported by teachers are displayed, including frequencies and proportions, excluding the segments coded as repetition. In the description of the results we only focus on the relevant codes. Figure 4.2 shows a frequency distribution of the main categories across conditions. In the name-only condition, teachers reported most cues from the Student\*Content category ( $M = 9.23$ ,  $SD = 5.90$ ). (Mis)conception was the most frequent code of all relevant codes in this condition ( $M = 3.48$ ,  $SD = 3.42$ ). In the name+answers condition, Student\*Content was also the most frequent main category ( $M = 14.46$ ,  $SD = 7.34$ ). In this condition, item performance was the most frequent code ( $M = 8.47$ ,  $SD = 5.70$ ). In the answers-only condition, Answers was the most frequent main category ( $M = 12.95$ ,  $SD = 6.55$ ). As in the name+answers condition, item performance was the most frequent code ( $M = 10.07$ ,  $SD = 5.58$ ). Although one might not expect student cues to be reported at all in the answers-only condition, teachers sometimes ( $M = 1.11$ ,  $SD = 1.62$ ) reported the "fabricated student" code (i.e., statements implying that teachers had an idea about the identity of the student or tried to guess the identity). Teachers occasionally also assigned characteristics to these fabricated students, and as a result also other student codes (e.g., "effort and work regulation") were reported sometimes in the answers-only condition.

**Table 4.4***Average Frequencies and Proportions of Codes and Main Categories per Student*

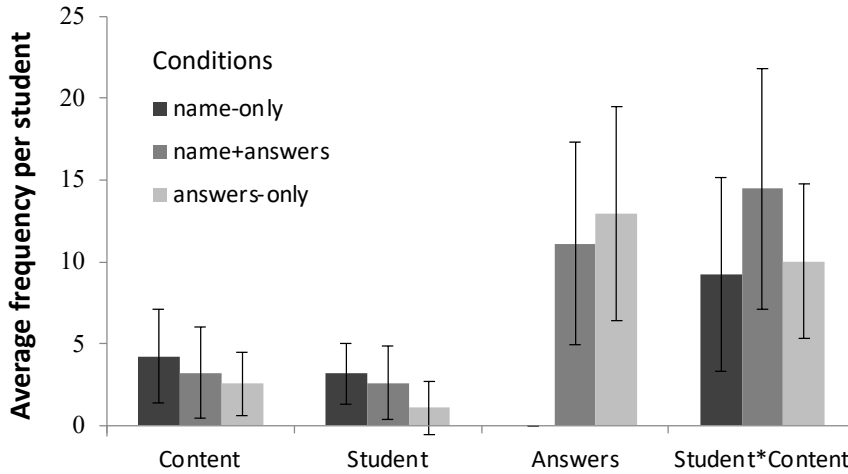
Assigned Codes	name-only		name+answers		answers-only	
	mean #	%	mean #	%	mean #	%
Relevant codes <sup>a</sup>						
Content						
Item characteristics	2.74	0.16	2.67	0.08	2.37	0.09
Curriculum	0.29	0.02	0.12	< 0.01	0.05	< 0.01
Instruction this lesson	1.19	0.07	0.44	0.01	0.14	0.01
<i>Total Content</i>	<i>4.23</i>	<i>0.25</i>	<i>3.23</i>	<i>0.10</i>	<i>2.56</i>	<i>0.10</i>
Student						
General cognitive	0.66	0.04	0.39	0.01	0.04	< 0.01
Math general	0.60	0.04	0.63	0.02	0.19	0.01
Other math domain	0.19	0.01	0.12	< 0.01	0.02	< 0.01
Effort and work regulation	1.00	0.06	0.65	0.02	0.26	0.01
Affective	0.40	0.02	0.40	0.01	0.09	< 0.01
Class behavior	0.02	< 0.01	0.02	< 0.01	0.00	0.00
Background	0.02	< 0.01	0.00	0.00	0.00	0.00
Gender	0.00	0.00	0.00	0.00	0.05	< 0.01
Student other	0.31	0.02	0.40	0.01	0.07	< 0.01
Fabricated student	0.00	0.00	0.00	0.00	0.39	0.01
<i>Total Student</i>	<i>3.19</i>	<i>0.19</i>	<i>2.61</i>	<i>0.08</i>	<i>1.11</i>	<i>0.04</i>
Answers						
Item performance	0.00	0.00	8.47	0.27	10.07	0.38
Overall test performance	0.00	0.00	2.63	0.08	2.88	0.11
<i>Total Answers</i>	<i>0.00</i>	<i>0.00</i>	<i>11.11</i>	<i>0.35</i>	<i>12.95</i>	<i>0.49</i>
Student*Content						
Understanding decimals	3.35	0.20	4.53	0.14	3.09	0.12
Strategy	1.95	0.12	1.42	0.05	0.35	0.01
(Mis)conception	3.48	0.21	7.72	0.25	5.98	0.22
Student guessed	0.24	0.01	0.35	0.01	0.35	0.01
Comparison other student	0.19	0.01	0.44	0.01	0.26	0.01
<i>Total Student*Content</i>	<i>9.23</i>	<i>0.55</i>	<i>14.46</i>	<i>0.46</i>	<i>10.04</i>	<i>0.38</i>
Total relevant codes	<i>16.64</i>		<i>31.40</i>		<i>26.65</i>	
Irrelevant codes <sup>b</sup>						
Affective teacher	0.21	0.01	0.86	0.01	0.39	0.01
Meta process teacher	1.65	0.04	2.12	0.04	1.98	0.04
Guessing	0.23	0.01	0.25	< 0.01	0.19	< 0.01
<i>Total Teacher</i>	<i>2.08</i>	<i>0.05</i>	<i>3.23</i>	<i>0.05</i>	<i>2.56</i>	<i>0.05</i>
Miscellaneous						
Judgment	15.06	0.38	15.65	0.26	14.96	0.29
Other	5.23	0.13	7.28	0.12	5.32	0.10
Unclear	0.98	0.02	1.68	0.03	1.56	0.03
<i>Total Miscellaneous</i>	<i>21.27</i>	<i>0.53</i>	<i>24.61</i>	<i>0.42</i>	<i>21.84</i>	<i>0.43</i>
<i>Total all codes</i>	<i>40.00</i>		<i>59.25</i>		<i>51.05</i>	

<sup>a</sup> Proportions reflect proportion of "total relevant codes".<sup>b</sup> Proportions reflect proportion of "total all codes". Repetitions are excluded.



**Figure 4.2**

*Effects of Cue Availability on the Frequency of Cues Reported by Teachers. Error Bars Indicate Standard Deviations*



#### 4.3.2.1. Cue Differences Across Conditions

The third Research Question was whether, and if so how, the cues reported by teachers differed across the conditions. Table 4.5 and Figure 4.2 display the frequency differences of the main categories across conditions. In describing the results, we focus on the significant differences. Content cues were reported most in the name-only condition, in which they were reported significantly more often than in the answers-only condition ( $B = 1.66$ ,  $d = .62$ ,  $p < .001$ ), but not significantly more than in the name+answers condition ( $B = 1.00$ ,  $d = .38$ ,  $p = .042$ ; not significant as Bonferroni correction for multiple hypotheses testing was applied, using an alpha level of  $.05/3 = .017$ ). This suggests that teachers use more cues related to curriculum content and material content when only student names are available, compared to when they have access to anonymized students' answers on practice problems.

As one would expect, student cues were reported significantly less frequently in the answers-only condition than in the name+answers condition ( $B = -2.09$ ,  $d = -1.00$ ,  $p < .001$ ) and the name-only condition ( $B = -1.51$ ,  $d = -.72$ ,  $p < .001$ ). Vice versa, as one would expect answer cues were reported significantly more in the answers-only ( $B = 12.95$ ,  $d = 1.86$ ,  $p < .001$ ) and name+answers ( $B = 11.11$ ,  $d = 1.44$ ,  $p < .001$ ) conditions than in the name-only condition, in which teachers did not have access to answer cues. Even though answer cues were reported most in the answers-only condition, this was not significantly more often than in the name+answers condition ( $B = 1.84$ ,  $d = .24$ ,  $p = .022$ ). Likewise, student cues were not reported significantly more often in the name-only condition than in the name+answers condition ( $B = 0.58$ ,  $d = .28$ ,  $p = .132$ ), suggesting that teachers did not rely less on their general knowledge about students when they had access to

students' practice answers in addition to their names.

Student\*Content cues were reported most in the name+answers condition and this differed significantly from both the name-only ( $B = 5.25, d = .81, p < .001$ ) and answers-only condition ( $B = 4.42, d = .68, p < .001$ ).

**Table 4.5**

*Parameter Estimates from a Multilevel Analysis on the Frequency of Assigned Codes*

Effects	B	SE	Cohen's <i>d</i>	<i>p</i>
Content				
Name-only vs. Name+Answers	1.00	0.50	0.38	.042
Name+answers vs. Answers-only	0.67	0.39	0.25	.086
Name-only vs. Answers-only	1.66	0.47	0.62	< .001*
Student				
Name-only vs. Name+Answers	0.58	0.38	0.28	.132
Name+answers vs. Answers-only	1.51	0.35	0.72	< .001*
Name-only vs. Answers-only	2.09	0.32	1.00	< .001*
Answers				
Name-only vs. Name+Answers	-11.11	1.22	-1.44	< .001*
Name+answers vs. Answers-only	-1.84	0.80	-0.24	.022
Name-only vs. Answers-only	-12.95	1.34	-1.68	< .001*
Student*Content				
Name-only vs. Name+Answers	-5.23	1.03	-0.81	< .001*
Name+answers vs. Answers-only	4.42	0.99	0.68	< .001*
Name-only vs. Answers-only	-0.81	0.91	-0.13	0.366

\* This effect significantly differed from zero when applying Bonferroni correction for multiple hypotheses testing, using an alpha level of  $0.05/3 = 0.017$

## 4.4. Discussion

The first aim of the present study was to investigate whether teachers' judgment accuracy of students' conceptual mathematics understanding would be affected by manipulating the availability of students' names, answers on a prior practice assignment, or both. This would lead to the (un)availability of certain cues on which teachers' judgments could be based. Teachers' judgment accuracy was measured by teachers' item-specific judgments of what students do understand (sensitivity) and judgments of what students do not understand (specificity) within the domain of decimal magnitude.

Our first hypothesis was that teachers' sensitivity and specificity would be higher when having access to students' answers on a practice assignment (considered to be predictive of students' actual understanding and therefore, of their performance on the test assignment) compared to when having access to only student names (which would result in activation of cues that we expected to have low predictive value). Our second hypothesis was that teachers' sensitivity and specificity would be higher when having access to only students' answers, compared to when having access to both students' answers and names. Contrary to our hypotheses, teachers' ability to indicate what students did understand (sensitivity) was not higher when students' answers on prior practice problems were available; it was even significantly lower when only students' answers were available, compared to when only names were available. Partly in line with the hypotheses regarding specificity, teachers were better able to judge accurately what students did not understand when they had access to their answers on prior practice problems, but only when they did not know who the students were (increase of .24 on a scale from 0 to 1). Although these findings show that the types of cues that are available to teachers may affect their judgment accuracy (mainly in terms of specificity), it does not tell us which cues teachers used exactly.

Therefore, the second aim of our study was to explore how teachers' cue use differed depending on the information types that were available (i.e., student cues, answer cues or both). The analyses of teachers' think-aloud data, recorded while they made judgments, showed that teachers in all conditions used cues related to the content of the task at hand and curriculum content. Not surprisingly, when teachers had only access to student names, they made no use of information on students' answers (which they did not have access to). Surprisingly, however, when teachers did not have access to student names, but only to students' answers, they still made some use of student cues (although significantly less than when student names were available). Teachers hypothesized, for instance, from which student the answers were ("fabricated student cues") and even assigned features to the anonymous students, such as having sloppy habits, having low concentration, being clever, or being uncertain. Another finding that we did not anticipate was that teachers also used cues that reflected an interaction between the student and the content material (e.g., statements referring to a decimal misconception held, or a strategy used by a student) of which it was mostly unclear whether these cues were derived from the student and/or the answers and/or the content material.

The differences in teachers' cue use across conditions can explain the differences in teachers' specificity, as we discuss in section 4.4.2. First, however, we discuss the findings regarding sensitivity.

#### 4.4.1. Effects of Cue Availability on Sensitivity of Teachers' Judgments

The finding that the sensitivity of teachers' judgments was higher when they had only students' names available compared to when they had only answer cues available was in contrast with our hypothesis (i.e., we expected that teachers' sensitivity would be higher when only answer cues were available). Rather than indicating that teachers made the most accurate item-specific judgments in the name-only condition, however, this finding probably reflects teachers' tendency to be more positive about their students' performance when they knew which student they were judging. Note that prior research, where the situations were comparable to our name-only condition, also showed that teachers generally overestimate their students (Artelt & Rausch, 2014; Klug et al., 2016). When teachers in the present study only knew the student's name, they judged, on average, almost 3 items more as having been answered correctly than when they only had students' answers on practice problems available. Given that students answered approximately two-thirds of the test items correctly, this means that when teachers would just randomly have assigned their "correct" judgments to the test items, the chance of judging a correct answer as correct was substantially higher in the name-only than in the answers-only condition.

#### 4.4.2. Effects of Cue Availability on Specificity of Teachers' Judgments

Specificity was higher, meaning that teachers were better able to accurately judge what students did *not* understand and would get wrong on the test assignment, when teachers had access to students' answers on prior practice problems, but only when they did not know who the students were. This finding may be explained by differences in teachers' cue use across conditions. We expected that having access to students' answers would result in more accurate judgments, because teachers would focus less on student cues and more on answer cues, the latter being presumably more predictive of students' actual understanding (see section 4.1.1.2). Indeed, teachers reported using significantly more answer cues when answers were available compared to when answers were unavailable (i.e., name-only condition), but they did not use *more* answer cues in the answers-only compared to the name+answers condition, so this cannot explain the higher specificity in the answers-only condition compared to the name+answers condition. Teachers also used fewer student cues when names were not available (i.e., in the answers-only condition), which is an unsurprising finding. More interestingly, however, having access to students' answers in addition to their names, did not result in the use of *fewer* student cues than in the name-only condition. Findings of Kaiser et al. (2015) already indicated that teachers' judgments of fictional students' mathematics grades were impaired when being provided with student characteristics in addition to information on students' oral and written

mathematics achievement. Our findings suggest that when teachers make judgments of their own students, focusing on student cues (triggered by access to student names) in addition to the answer cues may also interfere with adequately using the answer cues. The following quote of a teacher, taken from the name+answers condition in our study, illustrates that even though relevant cues (i.e., answer cues) were available, teachers may erroneously disqualify the relevant cues based on their knowledge about the student (i.e., student cues): "She places 0.13.... ah that's interesting, she places it behind the one [teacher analyzes a student's practice problem]. So then she thinks ... Well, that's sloppiness. I shouldn't take this one into account."

Another potential explanation for why specificity was higher in the answers-only condition than in the name-only condition might lie in differences in the use of content cues (i.e., cues related to curriculum content and material content). The findings show that teachers used significantly fewer content cues in the answers-only than in the name-only condition. According to Thiede et al. (2015), content cues are not predictive of students' actual understanding, leading to inaccurate teacher judgments. The same seems to apply to accuracy of students' own judgments: use of content cues led to less accurate judgments of their own text understanding (Thiede et al., 2010). In sum, making less use of student and content cues, and more use of answer cues, can explain why teachers' judgments of what students did not understand are most accurate when only having access to students' answers.

#### **4.4.3. Limitations and Future Research**

This study has several limitations. The decimal magnitude assignments included multiple-choice answers. The advantage of multiple choice was that it allowed us to construct the test in such a way that all potential misconceptions could be detected (not only the dominant ones). Unfortunately, this also meant that students could correctly guess answers. This may have led to the relatively high test scores (with students correctly answering approximately two-thirds of the test items), which in turn may have affected the sensitivity measure of judgment accuracy (see also section 4.4.1), and may explain why the answer cues in the present study had only modest predictive value for students' actual understanding. Another potential limitation is that item format (e.g., number lines vs. asking to circle a number) might affect predictive value. Because there was an imbalance between the number of number line and other tasks, we could not reliably examine judgment accuracy by task type in the present study. Future research could further investigate if the specific format of the items would influence their predictive value, even if they test the same conceptual knowledge.

Nevertheless, even though they had only modest predictive value, teachers made more accurate specificity judgments when having access to the answer cues compared to only student names or to both answers and names. Hence, if answer cues can be

defined in such a way that they have higher predictive value in future research, this can be expected to lead to even more accurate judgments and might provide teachers with useful tools that they (can) use in class when teaching mathematics (and other subjects) to monitor students' understanding and provide adaptive support.

Future research might as well consider including measures of teachers' knowledge of students' misconceptions, since more knowledge of misconceptions might lead to more accurate judgments (cf. Ostermann et al., 2017). Finally, our sample was relatively small, even with a within-subjects design, so including a larger sample of teachers and students, would be desirable in future research.

#### **4.4.4. Conclusions**

As prior research (Furnari et al., 2017; Hurwitz et al., 2007; Kaiser et al., 2013, 2015, 2017; Paleczek, 2017; Ready & Wright, 2011) indicated, teachers' knowledge of general student characteristics plays a major role in teachers' judgment processes. We examined how giving teachers access to students' answers on practice problems, additional to or instead of their general knowledge of specific students (triggered by access to students' names), affected teachers' judgment accuracy. The findings suggest that giving teachers access to the answers, in addition to knowledge of their students, does not make teachers focus less on student characteristics, and a result, does not significantly improve teachers' accuracy of students' decimal magnitude understanding. Giving teachers access to students' answers only (i.e., instead of their knowledge about students), seems to be especially effective for judging what a student does not yet understand. Our study shows that applying the cue-utilization approach in research on teachers' judgments may be a promising way to identify starting points for interventions for improving teachers' judgment accuracy, which ultimately may foster the quality of teachers' instructional decisions.



# Chapter 5

## Effects of Cue Availability on Primary School Teachers' Accuracy and Confidence in Their Judgments of Students' Mathematics Performance

This chapter was published as: Oudman, 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, 103982. <https://doi.org/10.1016/j.tate.2022.103982>

Author contributions: All authors designed the study. SO recruited participants, collected and analyzed the data, and drafted the manuscript. All authors contributed to critical revision of the manuscript. JvdP and TvG supervised the study.



## Abstract

We investigated how the accuracy of teachers' judgments of their students' performance on procedural mathematical tasks, as well as their confidence in that, can be improved. Thirty-three primary school teachers judged how their students ( $N = 553$ ) would perform on a multiplication and division task, with and without having access to performance cues (i.e., students' performance on similar tasks completed one week earlier). When available, teachers mostly seemed to base their judgments on performance cues. Availability of performance cues improved teachers' judgment accuracy, resulted in higher confidence in their judgment accuracy, and increased awareness of their judgment (in)accuracy.

## 5.1. Introduction

To optimally stimulate student learning, teachers need to provide ‘differentiated’ instruction; instruction that is adapted to students’ current level of performance or understanding (Parsons et al., 2018; Tomlinson et al., 2003; Van de Pol et al., 2010). In order to make adaptive instructional decisions, teachers’ judgments of their students’ performance or understanding—also called monitoring judgments—need to be accurate (see for empirical studies: Klug et al., 2013; Van de Pol et al., 2014; see for review studies Urhahne & Wijnia, 2021; Thiede et al., 2019). In the present study, we focus specifically on the accuracy of primary school teachers’ judgments of students’ mathematical performance, which are predominantly inaccurate and typically too optimistic (Gabriele et al., 2016; Oudman et al., 2018; Thiede et al., 2015, 2018, 2019; Zhu & Urhahne, 2018). This is problematic as it can result in instructional decisions that are too optimistic, and thus, not adapted to a student’s actual needs (Urhahne & Wijnia, 2021). For instance, teachers who overestimate their students’ performance may provide them with tasks that are too difficult or fail to provide additional instruction to those who need it. This results in suboptimal learning progress and increases the likelihood of students failing at subsequent tasks, which may have adverse motivational and emotional effects (Seegers & Boekaerts, 1993).

To encourage teachers to make accurate monitoring judgments and in turn adaptive instructional decisions, the importance of *Data Based Decision Making* (DBDM; Campbell & Levin, 2009; Schildkamp et al., 2017) and *formative assessment* (Black & Wiliam, 2009; Van der Kleij et al., 2015) is increasingly emphasized in educational policy. DBDM interventions focus mostly on using data at a ‘macro level’, such as standardized assessments that students complete a few times a year, to make educational decisions with instructional purposes but also school development and accountability purposes (Schildkamp et al., 2017). Formative assessment focuses more on the ‘micro level’, that is, on eliciting and using student performance data to inform instructional decisions in the classroom, both during and in between lessons (Black & Wiliam, 2009). In line with the latter, in the present study, we investigate whether providing teachers with students’ mathematical performance data at the micro level helps them more accurately predict students’ future task performance. Prior studies showed that providing teachers with information on students’ performance on a related prior task (which they can use in addition to their knowledge of students’ general characteristics, such as nationality or learning problems, when making judgments), does not necessarily lead to more accurate judgments (Oudman et al., 2018; Van de Pol et al., 2021). It is important to gain knowledge about which type of tasks, that are part of current mathematics education, would provide teachers with performance data that increase the accuracy of their judgments.

Teachers need to make accurate judgments of students' performance, and, in order to make effective follow-up decisions, they also need to be *aware* of their (in)accuracy (Gabriele et al., 2016). That is, when teachers are (rightfully) confident that their judgment is accurate, they will carry out appropriate instructional actions based on those judgments; when they are not confident that their judgment is accurate, they can first seek more information about a student's performance before taking instructional actions (Gabriele et al. 2016). Therefore, a second aim of the current study was to explore how manipulating the availability of performance data affects teachers' awareness of their judgment (in)accuracy.

Before introducing our research questions, we will first explain (1) how we define teachers' judgment accuracy and how this is influenced by the use of *cues* (i.e., specific pieces of information that can be used to inform judgments; Koriat, 1997, Cooksey et al., 2007), (2) how using certain cues and ignoring others can improve teachers' judgment accuracy, and (3) the concepts of teachers' *confidence in* and *awareness of* their judgment (in)accuracy.

### 5.1.1. Teachers' Judgment Accuracy and Cue Use

To assess the accuracy of teachers' judgments of their students' performance, different measures can be used (cf. Urhahne & Wijnia, 2021, for a discussion of different measures). Because our ultimate goal is to help teachers make more adaptive instructional decisions, we are mainly interested in *absolute accuracy*, defined as the absolute discrepancy between a teacher's judgment (i.e., prediction of how a student performs on a task) and student's actual performance on that task. For example, consider a teacher who expects that a student will correctly solve four out of ten problems, but the student only solves two out of ten problems correctly. In this case, the absolute deviation is two problems (4 minus 2) on a scale from zero to ten, with values closer to zero indicating higher accuracy. Prior research has also frequently used *bias*, that is, the signed version of absolute accuracy, indicating whether and how much teachers over- or underestimate their students' performance. Because overestimation and underestimation cancel each other out when averaging scores, this measure does not always reflect the extent to which judgments are actually accurate when using it in regression analyses. *Relative measures*, such as rank components, have also frequently been used in prior research and can be useful in educational contexts, for example to gain insight into which students are *most* in need of additional support. However, it is possible to make a perfect rank order in terms of students' performance, while, for instance, overestimating the actual performance of all students. When it comes to tailoring instructional activities to individual students' needs, for example, when deciding which students need additional instruction or which students are ready for a more difficult task (cf. mathematics lesson books such as Baak et al., 2018 and Borghouts et al., 2019, or EDI, a widely applied teaching model: Hollingsworth & Yabarra, 2018), teachers should be able to accurately judge students' actual task performance.

### 5.1.1.1. How Cue Use Affects Judgment Accuracy

Unfortunately, teachers' monitoring judgments of their students' performance or understanding are often inaccurate. An explanation for why this is the case can be found in Koriat's (1997) cue-utilization perspective on monitoring accuracy. According to this foundational theory, judgments are based on specific pieces of information (i.e., cues) that differ in the extent to which they are actually predictive of students' performance.

When judging how well a student will perform on a future test or task, teachers can use different types of cues (Thiede et al., 2019). In this study, we specifically focus on student cues and performance cues. *Student cues* are general characteristics of students. Student cues that are repeatedly reported in prior studies are students' general cognitive ability, nationality, SES, sex or gender, classroom engagement, conscientiousness, disability status, self-concept, and interest (e.g., Cooksey et al., 2007; Furnari et al., 2017; Gortazar et al., 2022; Helwig et al., 2001; Hurwitz et al., 2007; Johnston et al., 2019; Kaiser et al., 2013, 2015; Meissel et al., 2017; Oudman et al., 2018; Paleczek et al., 2017; Ready & Wright, 2011; Van de Pol et al., 2021; Zhu & Urhahne, 2020). *Performance cues* consist of information about students' prior performance on the same skills or content that the teacher is judging (cf. Van de Pol et al., 2021). For instance, teachers can decide which students do not yet master multi-digit multiplication (and need additional instruction) based on formative assessments such as, (1) students' scores on a task or test about the same problem type, completed one or multiple days earlier, or (2) how well students can answer practice problems during the whole-class instruction (Hollingsworth & Yabarra, 2018; Thiede et al., 2015; 2018).

According to Koriat's (1997) work on cue-utilization, teachers' judgments are more accurate when the cues being used are more *diagnostic*, that is, predictive of students' actual performance (cf. Thiede et al., 2019). The diagnosticity and use of cues can be graphically displayed by means of the Lens Model developed by Brunswik (1955; Figure 5.1), in which the analogy of a convex lens is used to display the relations between a judgment (in this study: judgment made by the teacher), cues, and the true state (in this study: student's performance). The teacher only "sees" the student's achievement (or other student related variables) through the "lens" of the cues (see Urhahne & Wijnia, 2021, for a review of literature applying the lens model to research on teaching). Thus, the key to improving teachers' judgment accuracy lies in fostering their use of more diagnostic cues.

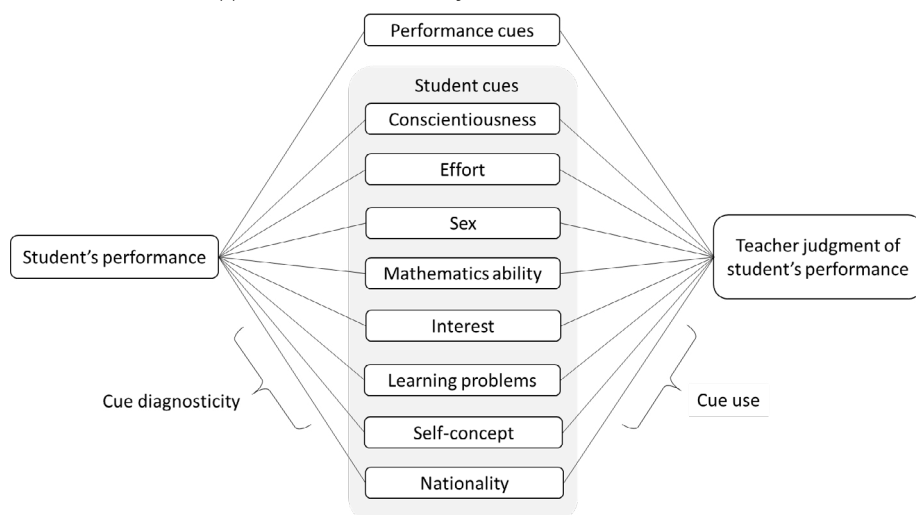
### 5.1.2. Improving Teachers' Judgment Accuracy: The Effect of Cue Availability

Prior studies suggest that performance cues are generally more diagnostic than student cues (in mathematics: Thiede et al., 2019; in text comprehension: Van de Pol et al., 2021). In many studies on teacher judgments, teachers had only student cues and no performance cues available (e.g., Furnari et al., 2017; Palecnek et al., 2017). Hence, it

is not surprising that their judgments were often inaccurate. Providing teachers with information from which they can derive performance cues, or focusing their attention on available information from which they can derive performance cues, might improve teachers' absolute judgment accuracy of their students' mathematical performance. Prior studies, however, did not find systematic evidence for this effect. In two studies, Thiede et al. investigated whether increased use of formative assessment practices—aimed at collecting performance cues to guide instructional decisions—improved primary school teachers' judgment accuracy with regard to mathematics. In one of the studies, Thiede et al. (2018) found that an intervention aimed at increasing teachers' use of formative assessment did not result in an improvement of teachers' relative accuracy and bias. In the other study, Thiede et al. (2019) found that the observed frequency with which teachers used formative assessment practices during their mathematics lessons was related to teachers' relative judgment accuracy, but not to their bias. Neither of those studies measured absolute accuracy.

**Figure 5.1**

*Brunswik's Lens Model Applied to the Current Study*



*Note.* Only cues included in the current study are displayed. The correlation between the cues and teacher judgment is an indication of teachers' cue use, the correlation between the cues and the student's actual performance is called cue diagnosticity.

Whereas the two studies of Thiede et al. focused on formative assessment practices in general, Zhu and Urhahne (2018) investigated the effects of using a specific tool which provided teachers with information about students' performance from which they could derive performance cues: Primary school teachers were asked to use learner response

systems (also referred to as clickers) approximately two times a week during their mathematics lessons. With the help of the learner response systems, teachers posed questions during whole-class instructions and received individual student's responses. This intervention improved teachers' relative judgment accuracy, bias, and absolute judgment accuracy: The absolute deviation between teachers' judgments and students' actual performance decreased from 10.43 to 4.14 items, on a test consisting of 25 items. Because the teachers in the Zhu and Urhahne (2018) study judged students' general mathematical skills (as was the case in the abovementioned studies of Thiede et al.), it remains unknown whether this intervention would also improve teachers' judgments of students' performance on specific mathematical tasks. This is important to establish because instructional decisions should be based on judgments of students' performance on the relevant tasks, and not on students' general mathematical performance (Baak et al., 2018; Borghouts et al., 2019; Hollingsworth & Yabarra, 2018; Stiggins & Chappuis, 2006). Moreover, as both the studies by Zhu and Urhahne (2018) and Thiede et al. (2018; 2019) did not provide information on the cues that teachers (presumably) used, it remains unclear whether the (lack of) increase in teachers' judgment accuracy in these studies was caused by (the absence of) improved cue use.

Studies that did measure teachers' cue use showed that giving teachers access to diagnostic performance cues did not necessarily improve their judgment accuracy; rather, the improvement seems to depend on the extent to which teachers simultaneously ignore less diagnostic student cues. For instance, Oudman et al. (2018) manipulated the type of cues primary school teachers had available when making judgments of their students' conceptual understanding of decimals. The teachers had access to: (1) only students' names (i.e., student cues), (2) only anonymized student work from which performance cues could be inferred, or (3) both students' names and their work (i.e., both student and performance cues). Teachers could infer students' decimal (mis)conceptions (i.e., performance cues) by analyzing students' work. The teachers thought aloud while making judgments, to measure their cue use. The findings suggest that teachers were most accurate when only performance cues were available (although this was only true for judgments of what students *did not understand*, not for judgments of what students *did understand*). When both student and performance cues were available, teachers did use performance cues, yet were not more accurate and did not focus less on student cues, compared to when only student cues were available. These findings suggest that it is hard to ignore student cues when both student and performance cues are available.

Similar results were found by Van de Pol et al. (2021). They investigated the relation between secondary school teachers' self-reported cue-utilization and their judgment accuracy of students' text comprehension. Using non-diagnostic student cues, such as effort and intelligence, in addition to diagnostic performance cues (characteristics of diagrams completed by the students, e.g., number of correct relations), appeared to

hamper teachers' absolute judgment accuracy. The findings also showed that teachers had difficulties with accurately inferring the performance cues: The teachers' judgment of the number of correct relations in a diagram completed by students deviated substantially from the actual number of correct relations in that diagram. When teachers' judgments of performance cues were inaccurate, their judgments of students' performance were also less accurate. Moreover, the findings indicated that regardless of whether the judgments of the performance cues were (in)accurate the teachers might have had difficulty with translating performance cues (e.g., number of correct causal relations in a diagram) into judgments (i.e., their estimates of the number of causal relations students would correctly recall on the posttest).

In summary, there may be three possible explanations for why providing teachers with performance cues does not always help improve their judgment accuracy: (1) when highly diagnostic performance cues are available, teachers do not merely use these performance cues but they also use less diagnostic student cues, (2) inferring performance cues from student work can be difficult, and (3) it is difficult to translate performance cues into judgments of students' performance. Therefore, the present study provided teachers with student work from which performance cues could easily be derived (as the scores are already provided) and that were highly aligned with the task teachers judged. In (Dutch) educational practice, teachers often have the opportunity to use this type of performance cue. For instance, students in the upper years of (Dutch) primary school work on basic procedural mathematics skills—addition, subtraction, multiplication, and division—on a weekly basis, sometimes as the main learning objective of a task and sometimes as part of a task with another main learning objective (Baak et al., 2018; Borghouts et al., 2019). These procedural mathematical tasks often contain performance information that is not difficult to infer: Tasks can unambiguously be scored in terms of number of problems answered (in)correctly. By means of formative assessment practices, teachers can elicit and use this information to infer cues that can inform their judgments and in turn, their instructional decisions. For example, over the course of two days, students work on a multiplication task. On the first day, they are introduced to the task and on the second day, they get to rehearse the task. Based on students' task scores on the first day, teachers could decide which students need additional instruction on the second day. Or, in the weeks after a monthly assessment, teachers could give additional instruction to students whose assessment performance (i.e., performance cues) indicate that they have not mastered a particular problem type.

We aim to investigate whether prompting teachers to use information from which performance cues can easily be inferred and are well aligned with the tasks that are to-be-judged, will positively affect their judgment accuracy. We also aim to investigate how prompting teachers to use this type of performance cue affects teachers' cue use. This might lead to insights on how to stimulate teachers to use more diagnostic cues, which

can in turn foster their judgment accuracy. Moreover, when using performance cues, it is unknown whether it is best for teachers to ignore *all* student cues or whether there are specific student cues that can be of added value. Even if the diagnosticity of student cues is generally lower than that of performance cues, it is possible that some student cues may have added diagnostic value when used in combination with performance cues. If certain student cues do indeed add diagnostic value, then using these could possibly lead to more accurate teacher judgments than merely using performance cues.

### 5.1.3. Teachers' Confidence in Their Judgment Accuracy

In order to make effective instructional decisions teachers not only need to make accurate judgments of their students' performance (e.g., Klug et al., 2013; Van de Pol et al., 2014), but they also need to be *aware* of their judgment (in)accuracy (Gabriele et al., 2016). Teachers show awareness of their judgment (in)accuracy when they feel relatively more confident about more accurate judgments and relatively less confident about less accurate judgments. This is typically measured by asking teachers how confident they are that their judgments are accurate, directly after making a performance judgment (some studies refer to this as second-order judgments: e.g., Dunlosky et al., 2005). When teachers are aware of their (in)accuracy they are likely to make appropriate instructional decisions, either based on judgments that were accurate and in which they have confidence, or by obtaining more information on the judgments that were inaccurate and in which they have less confidence. When teachers are *not aware* of their (in)accuracy, they either feel confident about less accurate judgments, which may lead to inappropriate instructional decisions, or they lack confidence in accurate judgments, which prompts teachers to seek more information (costing time and effort) when in fact this is not necessary (Gabriele et al., 2016). Particularly the combination of less accurate judgments and high confidence can have negative consequences because, in those cases, teachers' instructional decisions are less likely to be tailored to students' needs. To the best of our knowledge, with the exception of Gabriele et al. (2016), teachers' awareness of their (in)accuracy with regard to their students' academic performance has not yet been investigated. Moreover, the measure used by Gabriele et al. (2016) does not allow us to answer to what extent teachers are aware of their judgment (in)accuracy. Furthermore, it remains unknown how the availability of performance cues affects teachers' confidence in, or awareness of, their (in)accuracy.

As teachers' instructional decisions may become more effective when the match between their judgment accuracy and the confidence in their judgment accuracy is higher, increasing teachers' confidence in their judgment accuracy seems desirable in a situation in which judgments also become more accurate. Teachers might feel more confident of the accuracy of the judgments that are based on performance cues than those that are not because teachers might expect that performance cues have high diagnostic value. This is supported by findings of Zhu (2019) who investigated which cues



primary school teachers report being most important to base their judgments of students' achievement on. The teachers reported that, when judging students' achievement, relying on the performance of the last test (i.e., a performance cue) is three to five times more important than relying on grades of other subjects or test anxiety (i.e., student cues). However, teachers' perceived effectiveness of using formative assessment to inform their teaching varies and this affects their willingness to carry out formative assessments (for a review, see Yan et al., 2021). It could be that teachers with a less positive attitude towards formative assessment think that basing their instructional decisions on task-specific performance cues is not more effective than basing their decisions on other cues, such as students' general mathematics ability. For these teachers, making performance cues available might not increase their confidence in their judgment accuracy. An increase in confidence would be desirable if teachers' accuracy also increases when performance cues are available compared to when performance cues are not available (which we indeed expect, see section 5.1.2). However, an average increase in both teachers' judgment accuracy and confidence does not necessarily lead to increased *awareness* of their (in)accuracy. This is because awareness is defined as how well teachers can *distinguish* between their more accurate, and their less accurate, judgments in terms of confidence. It remains unknown how the *relation* between teachers' accuracy and their confidence—as indication of teachers' awareness of their (in)accuracy—changes after performance cues are made available to teachers.

#### 5.1.4. The Present Study

The present study aims to investigate 1) how prompting primary school teachers to use performance cues will affect their cue use when judging their students' performance on mathematical tasks and 2) how use of different (combinations of) cues affects their judgment accuracy, confidence in their judgment accuracy, and awareness of their judgment (in)accuracy. We specifically focus on procedural mathematical tasks (multiplication and division) that form a large part of the mathematics curriculum in the upper years of (Dutch) primary school (e.g., Baak et al., 2018; Borghouts et al., 2019). Teachers made judgments about how many of the six multiplication problems (e.g.,  $6 \times 472$ ) and six division problems ( $282 : 6$ ) their students answered correctly: These were made under two conditions, one with and one without having access to performance cues. Performance cues consisted of the number of problems students answered correctly on similar tasks (problems with same solution procedure and difficulty, but different numbers) completed one week earlier. Thus, these performance cues were well aligned with the to-be-judged-tasks and did not require interpretation of student work (as the number of problems answered correctly was given).

Our first research question (RQ1) was: How does availability of performance cues, that can easily be derived from student work and are well aligned with the task performance teachers have to judge, affect teachers' judgment accuracy of students' mathematical

task performance? We expected that when performance cues are available that teachers would make more accurate judgments, because (1) performance cues are generally diagnostic (Thiede et al., 2019; Van de Pol et al., 2021), (2) if the provided information does not require interpretation then inaccurate cue judgments might not appear and therefore do not hamper judgment accuracy, and (3) the more the cues are aligned with the to-be-made performance judgments, the easier it might be to translate the cue judgments into judgments of students' performance.

Second, we investigated how availability of performance cues affects teachers' use of student and performance cues, as indicated by the degree to which the cues predict teachers' judgments (RQ2). We would expect that teachers who are provided with information that does not require interpretation and is easy to translate into judgments, are less inclined to additionally use student cues when judging students' performance.

Third, in order to make accurate judgments, teachers should use highly diagnostic cues (Thiede et al., 2019; Van de Pol et al., 2021), and thus, we need to have knowledge of which cues are diagnostic for students' performance on procedural mathematical tasks. Therefore, we investigated to what extent student and performance cues predict primary school students' performance on procedural mathematical tasks, as an indication of cue diagnosticity (RQ3). In line with Van de Pol et al. (2021) we expected performance cues to be more diagnostic than student cues, although the present study is concerned with a different school subject and age group.

Fourth, assuming the performance cues in the present study are indeed more diagnostic than student cues, and teachers actually use the performance cues, leading to more accurate judgments, then it is relevant to find out whether it is best for teachers to ignore *all* student cues when performance cues are available, or whether there is potential added value in using some student cues in combination with performance cues (RQ4). In other words, we explored if there are student cues that, on top of performance cues, increase the explained variance in students' task performance. We had no specific expectations regarding this question, given a lack of prior studies on cue diagnosticity with regard to primary school procedural mathematical tasks.

Fifth, it has been suggested that teachers not only need to make accurate judgments of students' performance, but that they also need to be *aware* of their (in)accuracy (see section 5.1.3). We explored how the availability of performance cues affects (a) teachers' confidence in their judgment accuracy, and (b) teachers' awareness of their judgment (in) accuracy (RQ5). Regarding RQ5a, we explored whether and to what degree teachers feel more confident when performance cues are available, compared to when performance cues are not available. The reason being that teachers might expect performance cues to be highly diagnostic. Note that an increase in teachers' *confidence* in their judgment accuracy does not necessarily lead to an increase in *awareness* of their (in)accuracy. Hence, we had no expectations regarding RQ5b about teachers' awareness of their (in)accuracy.

## 5.2. Method

### 5.2.1. Context of the Present Study: Dutch (Mathematics) Education

In the Netherlands, students go to primary school from age 4 to age 12. The average class size at the time of writing is 23 students (Rijksoverheid [Dutch national government], n.d.). In the primary schools that participated in the current study, mathematics was taught by the teacher who also teaches most other subjects, as is the case in most Dutch primary schools. From age 6-7 onwards, students receive about one hour of formal mathematics education, daily. At age 7-8 students start with multiplication and around age 9 with division. From then on, multiplication and division are covered about weekly in the mathematics curriculum until the end of primary school, periodically becoming more complex. Sometimes multiplication and division are the main learning objective of a task, sometimes it is part of a task with another main learning objective (Baak et al., 2018; Borghouts et al., 2019; SLO, 2021).

### 5.2.2. Participants

#### 5.2.2.1. Teachers

Thirty-four teachers, teaching 9- to 10-year-old students (Dutch grade 6), volunteered to participate in this study. Teachers were recruited via (1) an advertisement on social media, (2) contacting schools that participated in a prior study (Oudman et al., 2018), and (3) the network of the first author who is also a primary school teacher. One teacher dropped out because they did not feel comfortable with completing the questionnaire about the student characteristics. The other 33 teachers (25 female) taught across 21 different primary schools in the Netherlands, ranged from ages 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 two to 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 5-9 months. Eight of the teachers had also been teaching their class in a previous grade.<sup>1</sup> This study received approval from the ethics review board of the authors' institute. The data of this study are openly available in an online depository at [https://osf.io/zv8en/?view\\_only=8cdea0c6a8314eac8cf6eb889307e628](https://osf.io/zv8en/?view_only=8cdea0c6a8314eac8cf6eb889307e628).

#### 5.2.2.2. Students

Of the 777 students who participated, data from 552 students were included in the analyses of the multiplication task and 553 in the analyses of the division task. Data

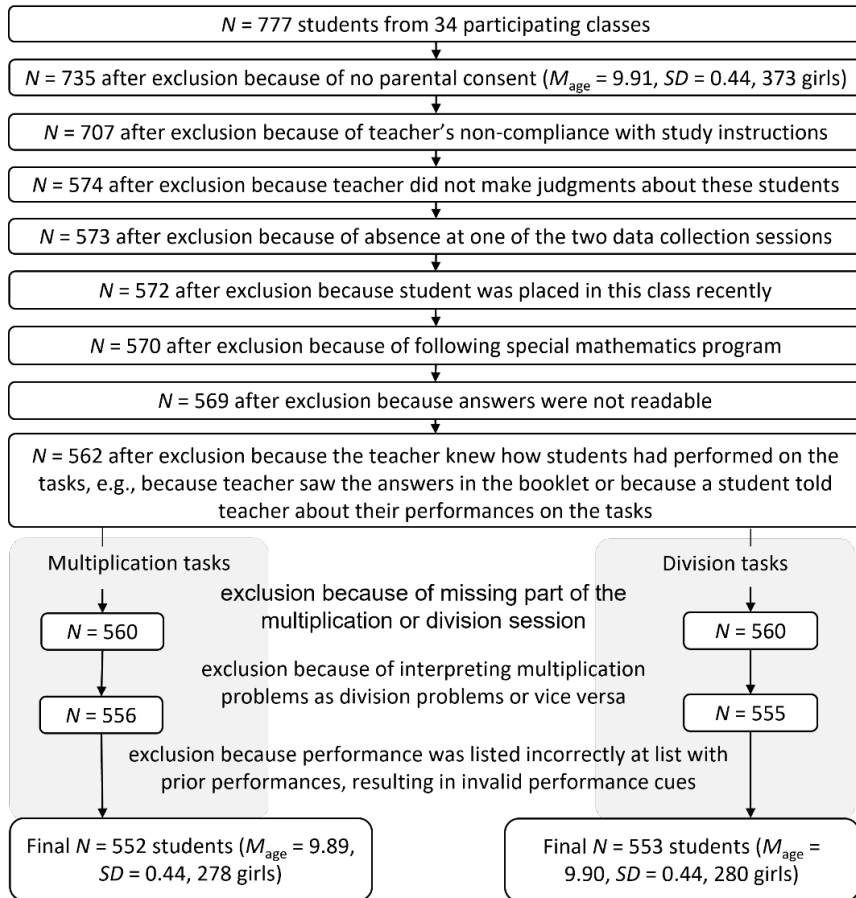
---

1 The eight teachers who taught their class also in a previous grade did not make significantly more accurate judgments than the other 25 teachers,  $p > .05$ .

from 545 students were included in both the analyses for multiplication and division. Figure 5.2 displays students' demographics and the number of students that had to be excluded and why.

**Figure 5.2**

*Flowchart of Reasons for, and Number of, Excluded Students*



*Note.* Excluded students were removed from the dataset. Multivariate outliers were defined for each analysis separately and are still included in the numbers in this flowchart.

### 5.2.3. Design

This study had a within-subjects design with two conditions: In the student-cue only condition, teachers made judgments and indicated their confidence in the accuracy of these judgments for 10 students, while being provided with the students' names (i.e.,

making student cues available). In the student+performance cue condition, teachers made judgments for 10 other students while being provided with the students' names and the students'

## 5.2.4. Materials and Measures

### 5.2.4.1. Students' Performance

On two days that were exactly one week apart, students made parallel versions (i.e., with isomorphic problems that have the same solution procedure and difficulty, but different numbers) of a multiplication and division task. On both days, students answered six multiplication problems (single-digit multiplicands multiplied by 3-digit multipliers, e.g.,  $6 \times 472$ ) and six division problems (3-digit dividends divided by single-digit divisors, e.g.,  $282 : 6$ ). Students received one point for each correctly answered problem, thus, per task performance scores ranged between 0 and 6. Students' task performance (i.e., how many problems they answered correctly) on day 1, were made available to teachers in the student+performance cue condition when making judgments on day 2.

### 5.2.4.2. Teachers' Judgment Accuracy

Per student, teachers were provided with the six multiplication or division items of day 2 and answered the question "How many of these six multiplication/division problems do you think this student answers correctly within 12 minutes?" on a 7-point scale ranging from 0 to 6. Based on students' task performance one week earlier, ten students per teacher were selected per condition, so that students with comparable scores were equally divided across the two conditions, within each class. This resulted in comparable means and variances of students' prior performance across conditions. When a class consisted of more than 20 students, we optimized the sample regarding the variability in student performance within each class (i.e., we avoided selecting students with similar scores as much as possible). When a class consisted of 20 students or less, teachers made judgments about all their students.

We analyzed teachers' *absolute accuracy*, determined by the absolute difference between the judged and actual performance (regardless of whether it was positive or negative), ranging from 0 to 6, with values closer to zero indicating more accurate judgments (Schraw, 2009; Urhahne & Wijnia, 2021). To allow for comparison of our findings to other studies using a different measure, we also report descriptive statistics of two other measures of teachers' judgment accuracy: Bias and the rank component. *Bias* was computed by subtracting students' actual performance from their judgment and ranges from -6 to 6, with values below zero indicating underestimation and values above

---

2 Students who have automated the procedures would need less than 10 minutes, based on the opinion of two mathematics experts and three experienced teachers, teaching 9-10 year olds.

zero indicating overestimation. The closer the values are to zero, the lower teachers' overestimation or underestimation of their student's performance is (Schraw, 2009; Urhahne & Wijnia, 2021). The *rank component* indicates how well teachers can accurately rank their students in terms of their performance. This measure was determined by the correlation between teachers' judgments and students' actual performance, thereby accounting for the non-independence of observations within classes, by applying multilevel regression models. The rank component ranges from 0 to 1, with values closer to 1 indicating higher accuracy (Schraw, 2009; Urhahne & Wijnia, 2021).

#### **5.2.4.3. Confidence Judgments**

Directly after teachers made a judgment of a student's performance, they made a confidence judgment by answering the question "How confident are you about the previous judgment?", on a 6-point Likert scale, ranging from "very unconfident" (1) to "very confident" (6).

#### **5.2.4.4. Cue Measures**

In the present study, performance cues consisted of students' performance on the multiplication and division task completed one week earlier, ranging from 0 to 6 per task. Student cues were measured using a teacher questionnaire, in which most cues were measured by one item and from the teachers' perspective: We assume that teachers use their own perception of student cues to base their judgments on (cf. studies that also used one-item teacher reports to measure teachers' cue use: Helwig et al., 2001; Kaiser et al., 2013; Zhu & Urhahne, 2020). The teacher-perceived student cues consisted of conscientiousness (during mathematics lessons), effort (during mathematics lessons), sex, interest in mathematics, general mathematics ability, nationality, presence of learning problems, and self-concept (students' confidence in their mathematical skills). See Figure 5.3 for an example item and section 5.1 of the Supplementary Materials for a list of student cue measures and the descriptive statistics per cue.

The data for this study were collected in the context of a larger project that also included student characteristics as perceived by teachers, for use in other studies. Students' intelligence was included in this 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 (.82). We additionally performed the analyses with intelligence instead of general mathematics ability as predictor and this led to similar conclusions as those reported here. Parents' educational level was also included in the questionnaire with the intention of use in the present study, but was removed from the analyses because teachers could not report this variable with certainty for most students.

**Figure 5.3**

*Example Item of Teacher Questionnaire About Student Cues.*

**This student works conscientiously during the normal mathematics lesson.**

*Examples: This student works orderly. This student works precisely.*

strongly disagree

disagree

agree

strongly agree

**5.2.4.5. Other Measures**

After making their judgments teachers indicated whether they had information about students' performance on the multiplication and division task on day 1 or 2, other than stemming from the performance cues they received (e.g., because they accidentally saw students' answers in a booklet or because students told them about their performance on the tasks). The reported students were deleted from the analyses (Figure 5.2).

After completing the student cue questionnaire, teachers answered a question about the extent to which tasks that were related to the multiplication and division tasks were part of their curriculum in the past week.<sup>3</sup>

**5.2.5. Procedure**

Data collection took place on two separate lesson days with exactly one week in between. On both days, the student and teacher session took place at the same time and lasted between 45 min and one hour. At least two weeks prior to the first day of the data collection, parents were informed and given the opportunity to object to their children's participation or use of their children's data.

**5.2.5.1. Students**

The student procedure was the same on day 1 and 2, except that isomorphic problems were used. After a short introduction by the experimenter, all students received a booklet and pen and completed the multiplication task, for which they had 12 minutes. It was emphasized that there was no need to hurry (as mentioned above, students who automated the procedures would need less than 10 minutes). When students finished the task in less than 12 minutes, they were instructed to read the (fiction) books they kept in their drawers. After 12 minutes, the experimenter gave the instruction that students who had not yet finished all problems should stop working. After the tasks, the students

---

3 This question was added because the multiplication and division tasks used in the present study were part of the regular curriculum. When these (or related) tasks would have been covered in the curriculum in the past week, teachers would have had more knowledge about students' multiplication and division skills (i.e., kind of performance cues) than when these (or related) tasks were not part of the past week curriculum. Adding this variable as predictor to the analyses did not change the significance of the results.

answered several questions<sup>4</sup> that were used for the larger project that the present study is part of. Finally, the same procedure was repeated for the division problems.

### **5.2.5.2. Teachers**

The teacher data collected on day 1 were not used in the present study, but in another study of the larger project (Oudman et al., 2022c). During the session on day 2, teachers were provided with a laptop, a list of names of students they had to make judgments about, noise-canceling headphones, and a covered list with the students' performance of one week earlier (i.e., performance cues). They sat in or close to their classroom so that they could not see their students working (as students were working on the tasks of day 2), yet would be able to intervene if an incident would occur in the classroom that required their attention (which was not the case). For each selected student, teachers completed a 1) multiplication performance judgment, 2) subsequent confidence judgment, 3) division performance judgment, and 4) subsequent confidence judgment. When teachers finished making judgments about the students in the student-cue only condition, they uncovered the list with students' prior performance and made the same four judgments for the students in the student+performance cue condition. After making the judgments, teachers indicated whether they had additional information about their students' performance on day 1 or 2 (see section 5.2.4.5). Next, they completed the questionnaire about student cues for students in both conditions. We assumed that teachers' perceptions of student characteristics would be influenced less by teachers' judgments of students' performance than teacher judgments by thinking about student characteristics. Finally, teachers answered the question about the past week's curriculum (see section 5.2.4.5).

### **5.2.6. Analyses**

All analyses for the multiplication and division task were performed separately because teachers' judgment accuracy could vary along with the subject matter (Kolovou et al., 2021). We performed multilevel regression analyses in Mplus version 8 (Muthén & Muthén, 1998-2017) to account for the nested data structure with students (level 1) clustered in classes, and thus in teachers as each teacher participated with one class (level 2). All fixed effects were tested at the student level. We used the maximum likelihood estimation with robust standard errors (MLR) which is robust to non-normality. Full output for all analyses, including intercepts and random effects, are presented in section 5.2. of the Supplementary Materials. To answer RQ1, about the effect of availability of performance cues on teachers' judgment accuracy, teachers' absolute accuracy was regressed on condition (student-cue

---

4 Students rated their invested effort, made a monitoring and regulation judgment (indicating their need for intervention regarding the type of problems they just completed, such as additional practice or instruction), and rated their feeling of confidence in the accuracy of the monitoring and regulation judgments they just made. The monitoring and regulation judgments were used in the study of Oudman et al. (2022a).



only condition, vs. student+performance cue condition).

RQ2 and 3 concern teachers' cue use and cue diagnosticity respectively and were analyzed in line with prior studies about how student characteristics relate to teacher judgments (e.g., Furnari et al., 2017; Meissel et al. 2017; Palecnek et al., 2017; Ready & Wright, 2011). RQ2 (cue use) was answered by regressing teacher judgments on the student cues as perceived by teachers. RQ3 (cue diagnosticity) was answered by regressing students' performance on the student cues as perceived by teachers. Table 5.2 (multiplication) and 5.3 (division) display the explained variance in teachers' judgments (in case of cue use) and in students' performance (in case of cue diagnosticity) by each cue, *including* shared explained variance by other cues we measured. The lens models in Figure 5.4 (multiplication) and 5.5 (division) display the explained variance in teachers' judgments and in students' performance by each cue, *excluding* shared explained variance by other cues we measured. Consequently, many coefficients representing diagnosticity and use are significant in Table 5.2 and 5.3, but not significant in the lens models (Figure 5.4 & 5.5), because the explained variance by these cues (almost) entirely overlaps with that of other cues. While we focus our analyses on the lens models, we do feel it is important to present the data in Tables 5.2 and 5.3 because these cues can still be diagnostic and teachers may still use these cues (although probably less than cues that are still significant in the lens models), even though the lens models might not give that impression.

To answer RQ4, about the potential added value of student cues for teachers' judgment accuracy, we compared the explained variance in students' performance by the performance cues only (i.e., students' prior performance on similar multiplication and division tasks) with the explained variance in students' performance by both the performance and student cues.

To answer RQ5a, about the effect of availability of performance cues on teachers' confidence in their judgment accuracy, we regressed teachers' confidence on condition (student-cue only condition, vs. student+performance cue condition). To answer RQ5b, about the effect of availability of performance cues on teachers' awareness of their (in) accuracy, we analyzed the effect of the interaction term between condition and teachers' accuracy on teachers' confidence in their accuracy. A significant interaction term would mean a stronger relation between teachers' accuracy and their confidence in one of the conditions, compared to the other condition, suggesting a difference in teachers' awareness of their (in)accuracy across conditions.

### **5.2.6.1. Missing Cases and Outliers**

When data were missing because students or teachers did not complete a question (this applied to 0.2-9.2% per variable), data were deleted list-wise in the analysis. For each multilevel model we analyzed, zero to 39 cases (a maximum of 7.1% of the data) were identified as multivariate outliers. We were mainly interested in the results without

outliers to avoid drawing conclusions that were potentially affected by extreme cases in our data. For the sake of transparency, we also ran the analyses with the inclusion of the outliers. This only led to a difference in statistical significance for the analysis of the multiplication task of RQ5b; thus, we reported both effects for this analysis, with and without outliers.

## 5.3. Results

### 5.3.1. Descriptive Statistics

Table 5.1 shows descriptive statistics for the performance, judgment, and confidence variables. In order to enable comparison with prior and future studies reporting on teachers' bias and the rank component, Table 5.1 also includes these two measures, in addition to teachers' absolute accuracy (the measure used in the analyses).

The intraclass correlation coefficient (ICC) for the main variables, reflecting the amount of between-teacher variability compared to the total amount of variability (both between and within teachers), were as follows: For teachers' absolute accuracy 2.5% for multiplication and 2.6% for division; For teachers' confidence in their accuracy 15.9% for multiplication and 13.2% for division. Thus, the largest part of the variability within the main variables resided at the within-teacher (i.e., student) level. For teachers' confidence in their accuracy, differences between teachers were more pronounced than for absolute accuracy.

### 5.3.2. Effect of Availability of Performance Cues on Teachers' Judgment Accuracy (RQ1)

In line with our hypothesis, teachers' judgments of students' multiplication and division performance were *more accurate* when both *student and performance cues* were provided than when only student cues were provided: See means of absolute accuracy in Table 5.1. The increase in accuracy was significant for both tasks (Multiplication:  $B = -0.43$ ,  $p \leq .001$ ; Division:  $B = -0.69$ ,  $p \leq .001$ ; Table S5.3, in section 5.2 of the Supplementary Materials). Teachers' judgment accuracy increased with 0.33 standard deviations for multiplication and with 0.52 standard deviations for division when making judgments with access to both student and performance cues, compared to only student cues. The effect size in terms of  $f^2$  is 0.03, indicating a small effect: 0.02 is the criterion for a small effect, 0.15 for a medium effect, 0.35 for a large effect (Cohen, 1988).

**Table 5.1**  
*Descriptive Statistics of Students' Performance, Teachers' Judgment and Confidence Variables*

Variable	Range	Multiplication			Division	
		Student cues only n = 279 <sup>a</sup>	Student+ performance cues n = 272 <sup>a</sup>	Student cues only n = 277 <sup>a</sup>	Student+ performance cues n = 275 <sup>a</sup>	
		M (SD)				
Student performance	0 to 6	3.76 (2.10)	3.88 (2.11)	2.94 (2.50)	3.07 (2.47)	
Judgment	0 to 6	3.69 (1.73)	3.95 (1.91)	2.96 (1.82)	3.04 (2.18)	
Absolute accuracy	0 to 6	1.58 (1.44)	1.14 (1.17)	1.77 (1.42)	1.08 (1.11)	
Bias	-6 to 6	-0.07 (2.23) <sup>b</sup>	0.08 (1.68) <sup>b</sup>	0.02 (2.31) <sup>b</sup>	-0.03 (1.59) <sup>b</sup>	
Rank component <sup>c</sup>	0 to 1	0.41 (0.08) <sup>***</sup>	0.65 (0.04) <sup>***</sup>	0.55 (0.05) <sup>***</sup>	0.78 (0.03) <sup>***</sup>	
Confidence in accuracy	1 to 6	4.31 (0.95)	4.69 (0.93)	4.06 (1.04)	4.66 (0.97)	
Accuracy awareness <sup>d</sup>	0 to 1	-0.17 (0.05) <sup>***</sup>	-0.17 (0.06) <sup>**</sup>	-0.21 (0.06) <sup>***</sup>	-0.38 (0.04) <sup>***</sup>	

<sup>a</sup> For the variables Absolute accuracy, Confidence in accuracy, and Accuracy awareness the samples sizes were somewhat smaller (max. 7% smaller), because multivariate outliers were removed to answer the research questions. <sup>b</sup> These values do not differ significantly from zero,  $p > .05$ . <sup>c</sup> Correlation between judgment and performance, accounted for the non-independence of observations within classes, followed by Standard Error between brackets. <sup>d</sup> Correlation between teachers' absolute accuracy and confidence in accuracy, accounted for the non-independence of observations within classes, followed by Standard Error between brackets. <sup>e</sup>  $p \leq .001$ , <sup>f</sup>  $p \leq .01$ , <sup>g</sup>  $p \leq .05$

**Table 5.2**

*Standardized Regression Coefficients, Indicating Cue Diagnosticity and Cue Use for the Multiplication Task, Including Shared Explained Variance by the Other Cues*

Cue	Diagnosticity $\beta$ (SE) <sup>a</sup>				Use $\beta$ (SE) <sup>b</sup>	
	Total sample	Student cues only	Student+ performance cues	Student cues only	Student+ performance cues	Student+ performance cues
Performance cues						
Multiplication	0.65 (0.03)***	0.63 (0.04)***	0.67 (0.04)***	-	0.93 (0.01)***	
Division	0.42 (0.04)***	0.45 (0.06)***	0.43 (0.06)***	-	0.54 (0.04)***	
Teacher-perceived student cues						
Conscientiousness	0.42 (0.04)***	0.42 (0.05)***	0.44 (0.05)***	0.43 (0.05)***	0.41 (0.06)***	
Effort	0.33 (0.04)***	0.40 (0.05)***	0.31 (0.04)***	0.47 (0.05)***	0.30 (0.05)***	
Sex (boy/girl) <sup>c</sup>	0.13 (0.05)**	0.11 (0.07)	0.18 (0.07)**	-0.19 (0.07)**	0.02 (0.08)	
Mathematics ability	0.41 (0.04)***	0.38 (0.06)***	0.46 (0.06)***	0.81 (0.03)***	0.65 (0.05)***	
Interest	0.36 (0.05)***	0.36 (0.07)***	0.38 (0.06)***	0.63 (0.05)***	0.49 (0.06)***	
Learning problems	-0.26 (0.05)***	-0.30 (0.08)***	-0.25 (0.06)***	-0.33 (0.07)***	-0.27 (0.07)***	
Self-concept	0.33 (0.05)***	0.33 (0.07)***	0.34 (0.06)***	0.61 (0.04)***	0.44 (0.06)***	
Nationality <sup>d</sup>	0.02 (0.04)	-0.02 (0.07)	0.09 (0.05)	-0.14 (0.07)*	-0.02 (0.07)	

<sup>a</sup>Standardized regression coefficients from multilevel regression models in which a single cue predicts students' performance. <sup>b</sup>Standardized regression coefficients from multilevel regression models in which a single cue predicts teachers' judgments. <sup>c</sup>This was an open question, but teachers only gave these two answers. <sup>d</sup>See for coding Table S5.2 in the Supplementary Materials. \*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

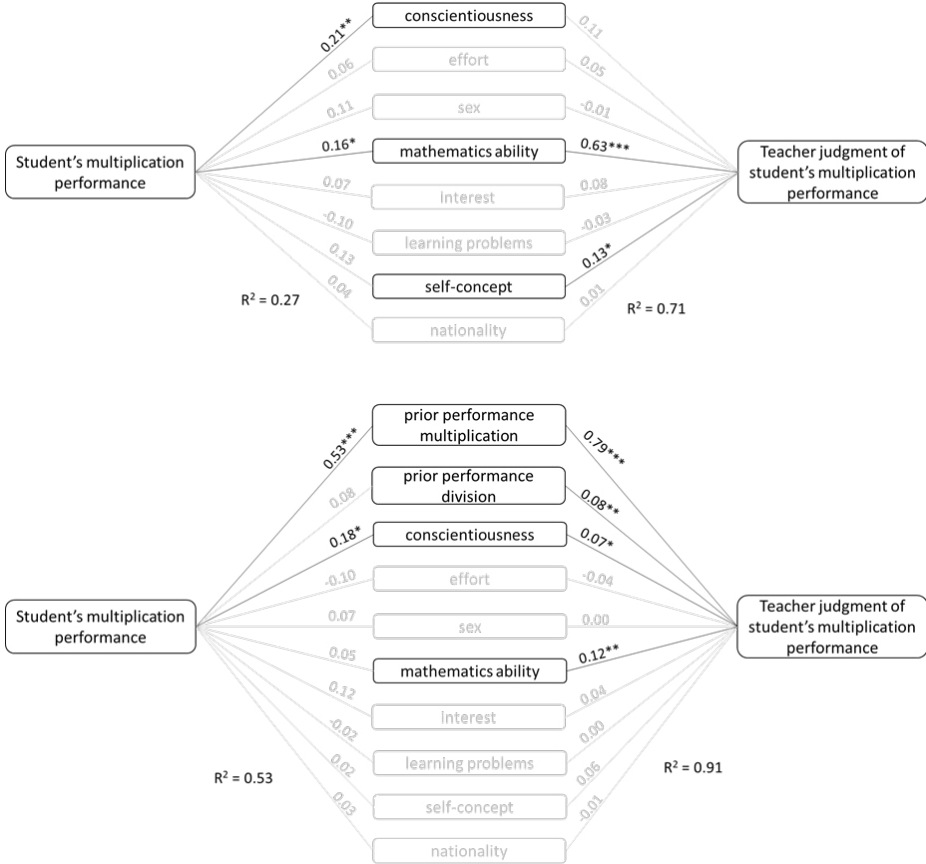
**Table 5.3**  
*Standardized Regression Coefficients, Indicating Cue Diagnosticity and Cue Use for the Division Task, Including Shared Explained Variance by the Other Cues*

Cue	Diagnosticity $\beta$ (SE) <sup>a</sup>			Use $\beta$ (SE) <sup>b</sup>		
	Total sample	Student cues only	Student+ performance cues	Student cues only	Student+ performance cues	Student+ performance cues
Performance cues						
Multiplication	0.41 (0.04)***	0.44 (0.06)***	0.46 (0.05)***	-	0.53 (0.05)***	
Division	0.77 (0.03)***	0.79 (0.03)***	0.79 (0.03)***	-	0.94 (0.01)***	
Teacher-perceived student cues						
Conscientiousness	0.28 (0.03)***	0.32 (0.04)***	0.27 (0.06)***	0.42 (0.06)***	0.30 (0.06)***	
Effort	0.34 (0.03)***	0.40 (0.06)***	0.32 (0.05)***	0.47 (0.05)***	0.32 (0.05)***	
SexSex (boy/girl) <sup>c</sup>	-0.04 (0.05)	-0.04 (0.08)	-0.01 (0.06)	-0.19 (0.07)**	-0.04 (0.07)	
Mathematics ability	0.53 (0.04)***	0.50 (0.05)***	0.59 (0.04)***	0.79 (0.03)***	0.68 (0.05)***	
Interest	0.44 (0.04)***	0.45 (0.05)***	0.45 (0.05)***	0.65 (0.05)***	0.50 (0.05)***	
Learning problems	-0.24 (0.04)***	-0.28 (0.06)***	-0.21 (0.05)***	-0.35 (0.07)***	-0.28 (0.06)***	
Self-concept	0.47 (0.04)***	0.43 (0.05)***	0.51 (0.05)***	0.63 (0.05)***	0.53 (0.05)***	
Nationality <sup>d</sup>	-0.06 (0.05)	-0.01 (0.06)	-0.05 (0.07)	-0.15 (0.06)*	-0.07 (0.06)	

Note. See Table 5.2 for explanations.

**Figure 5.4**

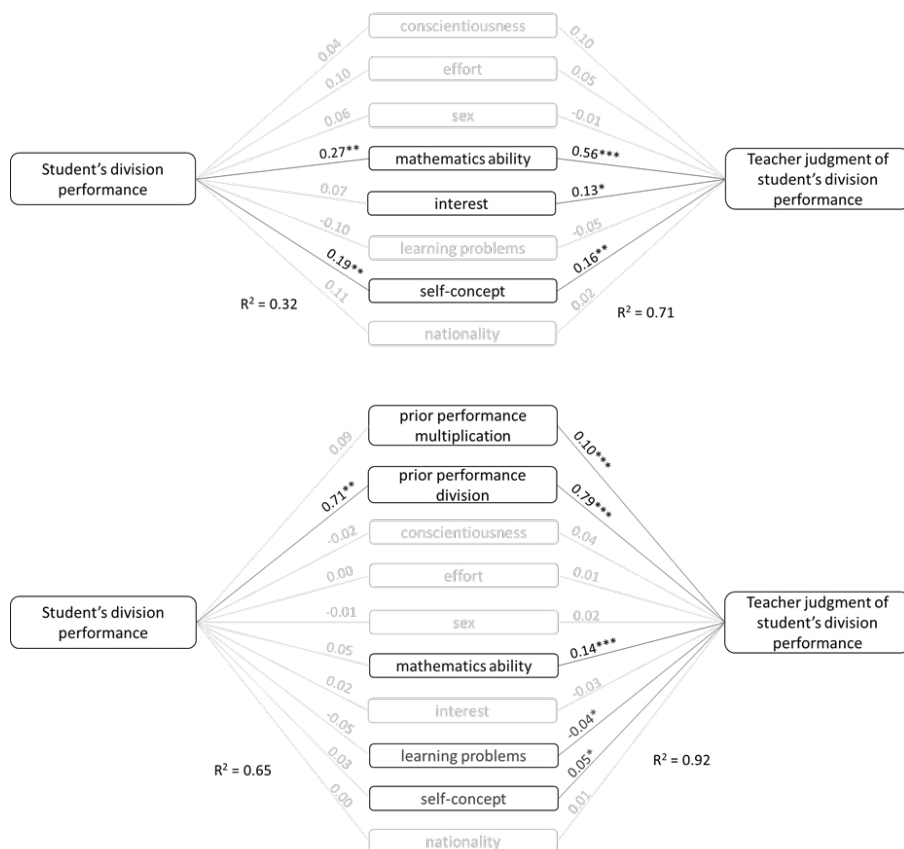
*Lens models of Teacher Judgments of Student's Multiplication Performance When Only Teacher-perceived Student Cues Were Available (Upper Model) and When Student and Performance Cues Were Available (Bottom Model)*



*Note.* Standardized regression coefficients on the left side of the model represent diagnosticity, standardized regression coefficients on the right side represent teachers' cue use.  $R^2$  is the explained variance in students' performance and teachers' judgments respectively, by all cues in the model. \*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

**Figure 5.5**

*Lens models of Teacher Judgments of Student's Division Performance When Only Teacher-perceived Student Cues Were Available (Upper Model) and When Student and Performance Cues Were Available (Bottom Model)*



Note. See Figure 5.4 for explanation.

### 5.3.3. Effect of Availability of Performance Cues on Teachers' Cue Use (RQ2)

As shown in the lens models (Figure 5.4 & 5.5), when *only student cues were available*, the uniquely explained variance in teachers' judgments by teacher-perceived *students' general mathematics ability* was the largest for both the multiplication and division task and at least three times larger than the variance explained by all other cues we measured. This suggests that when only student cues were available and teachers made judgments of students' performance, that they predominantly used their perceptions of students' general mathematics ability.

When *both student and performance cues were available*, the uniquely explained variance in teachers' judgments by *students' prior performance on a similar task* (i.e., performance cues) was the largest for both multiplication and division, and at least five times larger than the variance explained by all other cues we measured. This suggests that when both student and performance cues were available and teachers made judgments of students' performance, that they predominantly used performance cues (i.e., students' prior performance on the relevant task). It should be noted that teachers did not always copy students' prior performance when making their judgments, as the correlations between the performance cues and teachers' judgments were high but not perfect (0.93 for multiplication and 0.94 for division; Table 5.2 & 5.3). In summary, when performance cues are made available, teachers' cue use may shift from predominantly using their perception of students' general mathematics ability to predominantly using the provided performance cues (Figure 5.4 & 5.5).

In the student-cue only condition, all teacher-perceived student cues accounted for 71% of the variance in teachers' judgments for both multiplication and division; in the student+performance cue condition, all teacher-perceived student cues and performance cues accounted for 91% of the variance in teachers' judgments for multiplication and 92% for division (Figure 5.4 & 5.5). The effect size in terms of  $f^2$  are 2.45 in the student-cue only condition (both tasks) and 10.11 (multiplication) and 11.50 (division) in the student+performance condition, indicating exceptionally large effects (0.35 is the criterion for a large effect; Cohen, 1988). This suggests that the cues we measured give a fair indication of the cues teachers actually used, in both conditions.

### 5.3.4. Diagnosticity of Student and Performance Cues (RQ3)

When teachers *only had their perceptions of student cues available*, the uniquely explained variance in students' performance (representing diagnosticity) by teacher perceived *students' general mathematics ability* (for both tasks), *conscientiousness* (for multiplication), and *self-concept* (for division) were significant (Figure 5.4 & 5.5). When *both performance cues and teacher-perceived student cues were available*, only the diagnosticity of *students' prior performance on a similar task* (i.e., performance cues; for both multiplication and division) and the student cue *conscientiousness* (for multiplication) were significant. Hence, students' general mathematics ability has 'unique' diagnostic value when only teacher-perceived student cues are available, but not when performance cues are made available. The performance cues were at least three times more diagnostic than the teacher-perceived student cues. In the student-cue only condition, all teacher-perceived student cues accounted for 27% of the variance in students' performance for multiplication and 32% for division; in the student+performance cue condition, all teacher-perceived student cues and performance cues accounted for 53% of the variance in students' performance for multiplication and 65% for division (Figure 5.4 & 5.5). The effect size in terms of  $f^2$



are 0.37 (multiplication) and 0.47 (division) in the student-cue only condition and 1.13 (multiplication) and 1.86 (division) in the student+performance condition, indicating large effects (Cohen, 1988). This suggests that the cues we measured, together, fairly predict students' performance, especially in the student+performance cue condition.

### **5.3.5. Potential Added Value of Student Cues for Teachers' Judgment Accuracy (RQ4)**

In the student+performance cue condition, the explained variance in students' performance *by students' prior performance* on similar multiplication and division tasks was 48% for multiplication and 64% for division. When *adding the teacher-perceived student cues* as predictors of students' performance, the explained variance increased only 5% (i.e., to 53%) for multiplication and only 1% (i.e., to 65%) for division (Figure 5.4 & 5.5). The effect size in terms of  $f^2$  for this increase are 0.06 for multiplication, a small effect, and 0.01 for division, a trivial effect (Cohen, 1988).

### **5.3.6. Effect of Availability of Performance Cues on Teachers' Confidence in Their Judgment Accuracy (RQ5a)**

Teachers felt *significantly more confident* of their judgment accuracy for both the multiplication and division task when they had *access to performance cues*, than when they did not have access to performance cues (Multiplication:  $B = 0.37, p \leq .001$ ; Division:  $B = 0.59, p \leq .001$ ; Table S5.6, in section 5.2. of the Supplementary Materials). Their confidence on average increased from "somewhat confident" when only student cues were available to "confident" when performance cues were also available (Table 5.1).

When teachers made judgments with access to both teacher-perceived student cues and performance cues, their confidence increased with 0.39 standard deviations for multiplication and with 0.57 standard deviations for division, compared to when they only had access to teacher-perceived student cues. The effect sizes in terms of  $f^2$  are 0.05 for multiplication and 0.10 for division, indicating small effects (Cohen, 1988), and thus a small increase in confidence, when performance cues are made available.

### **5.3.7. Effect of Availability of Performance Cues on Teachers' Awareness of Their (In)Accuracy (RQ5b)**

The correlation between teachers' judgment accuracy and the confidence in their accuracy, analyzed separately per task and condition, were negative and significant (Table 5.1). This suggests that teachers were *aware of their (in)accuracy*, as they felt more confident of more accurate judgments than of less accurate judgments and vice versa. This was the case for both tasks and both conditions. When their accuracy increased with 1 point on the 7-point scale, their confidence increased between 0.10 and 0.30 on a 6-point scale (Table S5.7, in section 5.2 of the Supplementary Materials). The effect

sizes in terms of  $f^2$  are 0.03 (multiplication, both conditions; small effect), 0.05 (division, student cues only; small effect), and 0.17 (division, student+performance cues; medium effect; Cohen, 1988).

To test how the availability of performance cues (i.e., condition) affected teachers' accuracy awareness (RQ5b), we analyzed the effect of the interaction term between condition and teachers' accuracy on teachers' confidence in their accuracy. For both tasks, the relation between teachers' accuracy and their confidence was significantly stronger when teacher-perceived student cues and performance cues were available, compared to when only student cues were available (i.e., the interaction term was significant and negative;  $B_{multiplication} = -0.10, p = .048^5$ ;  $B_{division} = -0.16; p = .003$ ; Table S5.8, in section 5.2 of the Supplementary Materials). This suggests that teachers were on average somewhat *more aware of their (in)accuracy* when *both teacher-perceived student cues and performance cues were provided* compared to only student cues.

## 5.4. Discussion

The present study investigated (1) how prompting teachers to use performance cues affects primary school teachers' cue use when judging their students' performance on procedural mathematical tasks and (2) how the use of different (combinations of) cues affects their judgment accuracy, confidence in their judgment accuracy, and awareness of their judgment (in)accuracy.

### 5.4.1. Teachers' Judgment Accuracy and Cue Use (RQ1 and 2)

In line with our hypothesis, we showed that giving teachers access to performance cues—that can easily be derived from student work and are well aligned with the task performance teachers have to judge—in addition to their perceptions of student cues to which they always have access, positively affected teachers' absolute judgment accuracy of their students' mathematical task performance (RQ1). It should be noted that the effect was small, which can perhaps be explained in two ways. First, the teachers were already fairly accurate when they only had access to their perceptions of student cues: their judgments deviated 26% for multiplication and 30% for division from students' actual performance (Table 5.1). Second, the diagnosticity of the performance cues was high but not close to perfect; we return to this issue in section 5.4.4.

The finding that teachers' accuracy significantly increased when performance cues were provided differs from the results of two prior studies (Oudman et al., 2018; Van de Pol et al., 2021) which showed that giving teachers access to performance cues, in addition to student cues, did not necessarily lead to more accurate judgments. This difference

5 When outliers were still included in this sample, the interaction effect was not significant for the multiplication task:  $B = -0.05, p = .299$ .

in findings can presumably be explained by differences in the type of performance cues provided: In the present study, the provided information did not require interpretation (i.e., in prior studies, teachers had to interpret students' answers, whereas here, they were provided with the number of problems the students answered correctly on a similar prior task) and the tasks that the performance cues originated from were highly aligned with the to-be-judged-tasks (i.e., the earlier completed problems were isomorphic to the to-be-judged problems). This might have made it easier for teachers to use the performance cues—and ignore less diagnostic student cues—when making judgments about students' performance. This is supported by our findings regarding the second research question: When both teacher-perceived student cues and performance cues were available, the teachers hardly used student cues in addition to the performance cues.

The findings regarding the second research question also indicate that, when performance cues were not available, teachers seemed to predominantly base their judgments on their perception of students' general mathematics ability. Students' general mathematics ability can be seen as a more global proxy of performance cues and it might not be surprising that, when performance cues on similar tasks are not available, teachers seem to have the tendency to base their judgments on their knowledge of students' general performance in the relevant subject. However, teachers' perceptions of students' mathematics ability were not much more diagnostic than other (low diagnostic) student cues, and substantially less diagnostic than the performance cues. This stresses how important it is that teachers collect task specific performance cues to base their judgments on.

#### **5.4.2. Cue Diagnosticity (RQ3 and 4)**

That teachers in the present study seemed to mainly use performance cues when these were available seems a good decision, as the performance cues were much more diagnostic than the teacher-perceived student cues (including the student cue general mathematics ability; RQ3). Although we do not have data on this, it is possible that teachers knew that the performance cues would be more diagnostic than student cues. This would be in line with the finding of Zhu (2019) that teachers reported that, when making judgments of their students' achievement, relying on last test performance (i.e., a performance cue) is much more important than relying on grades of other subjects or test anxiety (i.e., student cues).

Even if student cues are less diagnostic than performance cues, it could have been possible that some teacher-perceived student cues would have had added diagnostic value when used in combination with performance cues. However, our findings with regard to the fourth research question suggest that use of teacher-perceived student cues (or at least the student cues we measured), in addition to using performance cues, would have little if any added value for the accuracy of teacher judgments.

### 5.4.3. Teachers' Confidence in and Awareness of Their (In) Accuracy (RQ5)

Finally, we explored teachers' awareness of their judgment (in)accuracy. This can be important for student learning, because when teachers are aware of their (in)accuracy they are more likely to make appropriate instructional decisions, either based on their judgments that were accurate, or by seeking more information about students' performance when their judgments were inaccurate (cf. Gabriele 2016). Teachers show accuracy awareness of their judgment (in)accuracy when they feel relatively more confident about more accurate judgments and relatively less confident about less accurate judgments.

As we expected, teachers felt more confident of their judgment accuracy when performance cues were available, than when only teacher-perceived student cues were available (RQ5a), and rightly so, as their accuracy was also higher. Interestingly, teachers' confidence in their accuracy only increased slightly and on average came close to 'confident' but not 'very confident'. Teachers might have known that the performance cues were more diagnostic than student cues, but also that the diagnosticity of the performance cues was not close to perfect.

An average increase in teachers' *confidence* in their judgment accuracy and in their *accuracy* does not necessarily lead to an increase in teachers' *awareness* of their (in) accuracy. The present study was the first to explore whether or not teachers were aware of their judgment (in)accuracy with regard to students' performance, and found positive results: Teachers indeed showed some awareness of their (in)accuracy, for both the multiplication and division task and when performance cues were or were not available. Moreover, teachers' accuracy awareness was positively affected by the availability of performance cues (RQ5b): Teachers were somewhat more aware of their (in)accuracy when performance and teacher-perceived student cues were available, compared to when only student cues were available. This finding could mean that when teachers use diagnostic performance cues, teachers' instructional decisions about procedural mathematical tasks are not only more accurate, but also more effective. For example, when teachers use performance cues instead of only their perceptions of student cues, they could be more likely to act upon accurate judgments that they are more confident about, and more likely to seek additional information when their judgments are inaccurate (Gabriele et al., 2016). Of course, this should be confirmed by future research.

### 5.4.4. Limitations and Future Research

One limitation of this study is that we did not directly measure teachers' cue use, but did so by means of correlations between teachers' judgments and measures of (teacher-perceived) cue values. The explained variance in teachers' judgments by the cues we measured was high (above .70 when only student cues were available and above .90 when student and performance cues were available). This suggests that the cues we measured

give a fair indication of the cues teachers actually used. As we did not directly measure cue use, it is possible that teachers did not actually use the cues that are indicated by our findings, but instead used cues that are related (both conceptually and correlational) to the cues we measured. For instance, as mentioned in section 5.2.4.4., teachers' perceptions of their students' intelligence and mathematics ability are highly related, so concluding which of the two variables they actually use is not possible via correlational research. Future research investigating the effect of interventions on teachers' cue use and their accuracy can include more direct measures of teachers' cue use via think aloud protocols or questionnaires about which cues they used. As teachers' perceptions of the student cues can differ from the actual cue values, for instance as measured by student questionnaires (Van de Pol et al., 2021), future studies could also investigate whether teachers' judgment accuracy would improve when teachers would more accurately judge the student cues. Additionally, these future studies could incorporate a check on whether the order of measuring judgments and cue use matters: We measured teachers' perceptions of cues after teachers made the judgments, but would the findings change when it was measured in the opposite order?

A question raised by our findings is how the diagnosticity of performance cues can be further improved. In the current study, the diagnosticity of the performance cues ranged between 0.63 and 0.79 (including shared explained variance by other cues), which is fairly high, but not close to perfect. If the diagnosticity of performance cues can be increased, the accuracy of teachers' judgments might also improve further. It would be interesting in future research to look for factors that influence the diagnosticity of performance cues, such as the type of task or the time between the task on which performance cues were collected and the to-be-judged tasks (i.e., a week in the present study). Future research could also attempt to measure to what extent cue diagnosticity differs across students and whether this can be explained by specific student characteristics. For instance, in theory, cues like effort or interest might be diagnostic when they are high or low in a student, but less diagnostic when they are medium/moderate.

Another question we cannot answer based on our data, is to what extent teachers were aware of the diagnosticity of the cues they used. As discussed earlier, our findings that teachers mostly used performance cues when these were available and hardly used student cues, might suggest that they were aware of the higher diagnosticity of performance cues. However, we do not know this for certain. Teachers' beliefs about and awareness of cue diagnosticity might influence their judgment accuracy and confidence in their accuracy. It would be valuable in future research to interview teachers about their thoughts on cue diagnosticity and how this affects their cue use, judgments, and confidence in their accuracy.

Finally, an important question is to what extent our findings would generalize. First, our findings apply to performance cues that can easily be inferred from student work and

are well aligned with the to-be-judged-tasks. Of course, even when teaching procedural tasks as the ones in the present study, teachers can also engage in interpreting students' strategy use to inform their instructional decisions. However, teachers often lack time (Schildkamp et al., 2017) and it is very time-consuming to analyze students' strategy use each time a student does not master a task. Using students' task scores to make quick decisions on which students do and do not master a task is an efficient method that can be alternated with more in-depth analyses of students' strategy use. While teachers commonly have access to 'quick' cues, these cues are not always available, for example when starting with a task that is new to the students. Future research could investigate for different type of tasks that are used in educational practice, and that vary to the degree to which they require interpretation, (1) whether they contain diagnostic performance cues, (2) whether teachers are able to use these performance cues and ignore less diagnostic cues, and (3) how this affects teachers' judgment accuracy and accuracy awareness. Future studies should also investigate further whether teachers can be trained to use performance cues that are more difficult to interpret (cf. those used in prior research; Oudman et al., 2018; Van de Pol et al., 2021) and at the same time ignore student cues. For instance, by asking teachers to make judgments based on vignettes after which they receive feedback about their use of student and performance cues. Lastly, it is an open question whether our findings would replicate in a larger sample of teachers and schools and whether they generalize to other age groups within and beyond primary school.

#### 5.4.5. Conclusions and Practical Implications

Formative assessment does not necessarily lead to accurate teacher judgments (Thiede et al., 2018, 2019) and this might be caused by the type of performance cues that are used by the teachers. Our findings indicate that the use of diagnostic performance cues that can easily be inferred from student work and are highly aligned with the task performance teachers have to judge (e.g., students' prior performance on similar tasks) improves teachers' judgment accuracy. It might seem obvious that teachers are able to use this type of performance cue and that this increases their judgment accuracy, but this was not a given when looking at prior research (Oudman et al., 2018; Van de Pol et al., 2021). Moreover, it was unknown whether teachers would also *ignore* less diagnostic student cues (as they continued to use those in prior studies), which indeed seemed to be the case. Furthermore, the present study showed that teachers are already somewhat aware of their (in)accuracy, in that they feel relatively more confident about more accurate judgments and relatively less confident about less accurate judgments, and that teachers' confidence in and awareness of their (in)accuracy can be positively affected by using performance cues, all of which is important for adaptive teaching.

Our findings suggest that encouraging teachers to use short formative assessment practices, which are relatively easy to implement, might help them to more accurately

evaluate their students' performance and needs. For instance, ending a mathematics lesson by asking students to solve a problem that represents the main learning objective of that lesson, and show their answers on mini-whiteboards they hold up (Wiliam, 2011), might provide teachers with the kind of easy to interpret performance cues that were also used in the present study, and help them to get a quick overview of which students might need additional interventions. When working with online learning systems, teachers can be encouraged to base their instructional decisions on the information on students' performance as shown in the teacher dashboards (which teachers do not necessarily consult; Molenaar & Knoop-van Campen, 2017). To help design and implement the most effective interventions, future research should test which tasks within the current (mathematics) curriculum do and do not provide performance cues that increase teachers' accuracy. This knowledge can then be included in teacher professional development programs aimed at improving formative assessment practices and teachers' judgment accuracy.







# Chapter 6

## Primary School Teachers' Judgments of Their Students' Monitoring and Regulation Skills

This chapter is submitted for publication as: Oudman, S., Van de Pol, J., Van Loon, M., & Van Gog, T. (2022). *Primary school teachers' judgments of their students' monitoring and regulation skills* [Manuscript submitted for Publication].

Author contributions: SO, JvdP, MvL TvG designed the study. SO recruited participants, collected and analyzed the data, and drafted the manuscript. All authors contributed to critical revision of the manuscript. JvdP and TvG supervised the study.

## 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 well primary school teachers can judge their students' monitoring and regulation accuracy and whether and how student characteristics influence this process. Thirty-three teachers, teaching 9- to 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 (before self-scoring) and regulation (before and after self-scoring) skills. 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 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' regulation accuracy before self-scoring seemed somewhat biased by students' nationality. These findings and measures can ultimately contribute to the design of interventions to help teachers judge and develop their students' self-regulated learning skills.

## 6.1. 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). In most models of self-regulated learning (see Panadero, 2017), two central processes are monitoring (evaluating one's performance) and regulation (controlling one's study activities; De Bruin & Van Gog, 2012; Griffin et al., 2013). Unfortunately, primary school students' self-monitoring and self-regulation are often inaccurate (e.g., Baars et al., 2014a; Oudman et al., 2022b; Prinz et al., 2020; Van Loon & Roebers, 2017). Prior studies have found interventions, such as asking students to self-score their work with the use of standards, to improve students' monitoring and regulation accuracy (in mathematics: Oudman et al., 2022b; in text comprehension: Van Loon & Roebers, 2017). However, especially students' regulation is 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). 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 influence these judgments.

### 6.1.1. Student Monitoring and Regulation Judgments

Accurate monitoring and regulation are necessary for effective self-regulated learning (Dunlosky & Rawson, 2012). Self-regulated learning theories generally assume that students' monitoring judgments influence their regulation judgments, and that accurate monitoring is a necessary (though not sufficient) precondition for accurate regulation (Dunlosky & Rawson, 2012; Metcalfe & Finn, 2008; Pintrich, 2000; Winne & Hadwin 1998;

Zimmerman, 2000). 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., 2022b) they will spend time on activities they already mastered rather than on those they need to learn.

In the context of primary school mathematics education, monitoring judgments are, for example, students' evaluations of their performance on a mathematical task they just completed. Common regulatory actions (in Dutch primary schools) when students do not yet master specific mathematical skills 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., 2014a; Dunlosky & Rawson, 2012; Oudman et al., 2022b). 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., 2022b).

### **6.1.1.1. 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., 2022b; 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., 2022b) 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.

### 6.1.2. 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 varies 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 (2022) 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, the teacher's judgment deviated 3.44 on a 24-point scale 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.

### **6.1.3. (Potential) Effects of Student Characteristics on Teacher Judgments of Students' Monitoring and Regulation Skills**

When making judgments about their students' *performance*, teachers seem to use their perceptions of general student characteristics (also called student cues) when making these judgments. 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., 2023), 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 that teachers use their perceptions of students' achievement level and self-concept as indicators of students' metacognitive abilities. Correlational analyses by Friedrich et al. (2013) suggest that teachers use their perceptions of students' mathematics competence as an indication 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 also affect 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.

#### **6.1.4. 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 influence 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 context of text comprehension, Van de Pol and Oudman (2022) found that secondary school teachers' judgments deviated 3.44 on a 24-point scale from students' actual monitoring accuracy. 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 might



use their perceptions of students' mathematics abilities, variables related to students' working behavior (such as effort and conscientiousness), and students' self-concept when making 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 6.1.3). However, because of a lack of prior research, we had no specific hypotheses.

## 6.2. Method

This study is based on a dataset of a larger project that also focuses on student monitoring and regulation judgments (Oudman et al., 2022a, 2022b) and teacher judgments of their students' performance (Oudman et al., 2023) in the context of mathematics problem-solving in primary school. Therefore, not all measures mentioned in section 6.3.2 and 6.3.3 were used in the present study, and there may be some overlap in the description of the method section with other papers.

### 6.2.1. Participants

#### 6.2.1.1. Teachers

Thirty-four teachers, teaching 9- to 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 5-9 months. Eight of the teachers had also been teaching their class in a previous grade.<sup>1</sup> This study received approval from the ethics review board of the authors' institute.

---

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,  $p > .05$ .

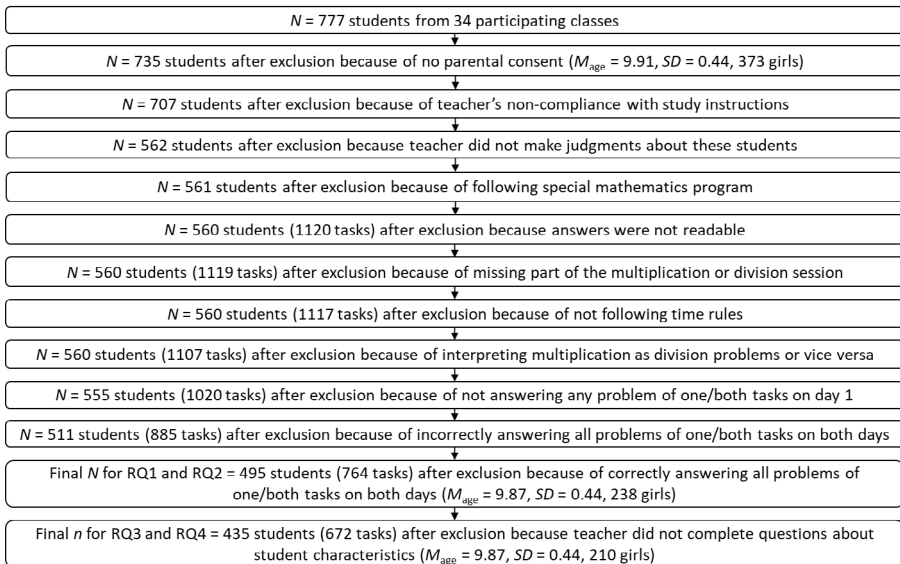
### 6.2.1.2. Students

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 6.3.2.2). Each student completed a multiplication and division task, but for some students, data from only one of the tasks could be used. Figure 6.1 displays students' demographics and the number of students and tasks that were excluded, including the reason for exclusion. As Figure 6.1 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.<sup>2</sup> 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., 2022b).

**Figure 6.1**

*Flowchart of Reasons for and Number of Excluded Students and Tasks*



*Note.* For some students, data from only one of the tasks was used. Missing cases and multivariate outliers were defined after exclusion of students and for each analysis separately (and are thus still included in this flowchart).

2 Similar tasks were administered on two days, but in the present study we only used student data of day 1, see Section 6.3.2.

## 6.2.2. 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 minutes and one hour.

### 6.2.2.1. 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.,  $6 \times 472$ ). They had 12 minutes to complete the task, but it was emphasized that there was no need to hurry.

When students finished the task in less than 12 minutes, they were instructed to read the (fiction) books they kept in their drawers. After 12 minutes, 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 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).

### 6.2.2.2. 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 minutes?". 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 Figure 6.2. 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 Figure 6.2. 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).<sup>3</sup> 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 S6.1 in the Supplementary Materials contains a list of the student characteristic measures, answer scales, and descriptive statistics per characteristic.

**Figure 6.2**

*Measures of Teacher Judgment of Student Monitoring/Regulation Judgment (TJSMJ & TJSRJ)*

What do you think this student answers here (before self-scoring)?

0 1 2 3 4 5 6

How many of the 6 **multiplication** problems do you think you solved correctly?

0 1 2 3 4 5 6

What do you think this student answers here (before self-scoring)?

I need additional instruction for this multiplication task.

I need additional practice with this multiplication task.

I need additional instruction **and** I need additional practice with this multiplication task.

I do **not** need additional instruction **and** I do **not** need additional practice with this multiplication task.

If you think about the 6 **multiplication** problems you just completed, what suits you best?

I need additional instruction for these multiplication problems.

I need additional practice with these multiplication problems .

I need additional instruction **and** I need additional practice with these multiplication problems.

I do **not** need additional instruction **and** I do **not** need additional practice with these multiplication problems.

### 6.2.3. 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 (2022) that builds on the literature about teacher judgments of their students' performance. Figure 6.3 (monitoring) and 6.4 (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 6.3.2).

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 (.82). Parents' educational level was also removed from the analyses because teachers could not report this variable with certainty for most students.

### 6.2.3.1. 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 Figure 6.3, 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., 2022b). 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 minutes (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 Figure 6.4 did not need additional instruction or practice, represented by the value ‘0’.

**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 Figure 6.3 thought they scored one point.

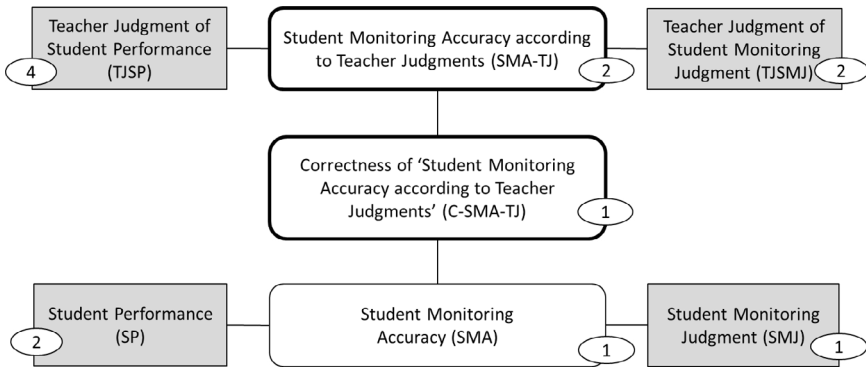
**Student Regulation Judgment (SRJ) Before and After Self-Scoring.** Students’ regulation judgments were coded as follows: (0) nothing needed, (1) additional practice needed, and (2) additional instruction needed (and practice afterwards). The needs ‘additional instruction’ and ‘additional practice and additional instruction’ were merged (for explanation see Oudman et al., 2022b). The student in the numeric example in Figure 6.4 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 Figure 6.3, 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 Figure 6.4, 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.

**Figure 6.3**

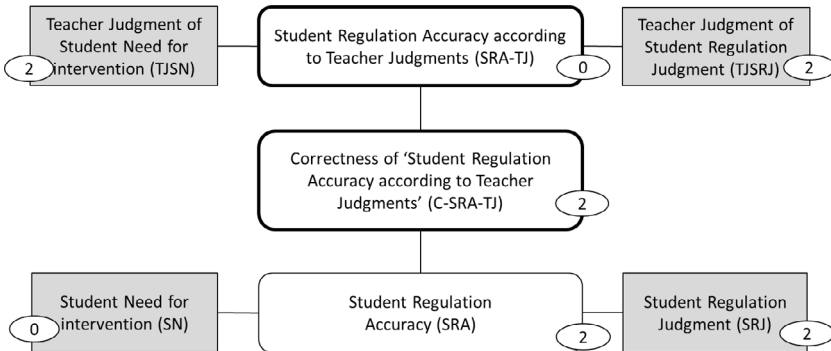
*Measurement Framework of Teacher Judgments of Students' Monitoring Accuracy*



*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.

**Figure 6.4**

*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.

### 6.2.3.2. 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 Figure 6.3 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 Figure 6.4 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 Figure 6.3 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 Figure 6.4, 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 Figure 6.3, 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 (TJSP & TJSMJ) and indicates to what degree the teacher thinks that the student makes an accurate judgment of their performance. In the numeric example of Figure 6.3, 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 (TJR). In the numeric example in Figure 6.4, 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 Figure 6.3, 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 Figure 6.4, 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.

#### 6.2.4. 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). We used the maximum likelihood estimation with robust standard errors (MLR) which is robust to non-normality.

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 different procedural mathematics tasks with regard to the role of student characteristics in 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% (see Table 6.1).

When data were missing because students or teachers did not complete a question (this applied to 0-9.7% per variable), data were deleted list-wise in the analysis. Per multilevel multiple regression model, which we performed to answer RQ3 and 4, zero to 0.7 % 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 additionally ran the analyses with outliers still included. This did not lead to different results.

The data of this study are openly available in an online depository at [https://osf.io/wh9r8/?view\\_only=f2a1ba6a0b5748efa366735adb747450](https://osf.io/wh9r8/?view_only=f2a1ba6a0b5748efa366735adb747450).

**Table 6.1***Intraclass Correlation Coefficients (ICC) of Main Study Variables*

		<b>Student Accuracy according to Teacher Judgments</b>	<b>Correctness of 'Student Accuracy according to Teacher Judgments'</b>
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).

## 6.3. Results

Descriptive statistics of all performance, need, and judgment variables are displayed in Table 6.2.

### 6.3.1. Teacher Judgments of Students' Monitoring Accuracy (RQ1)

#### 6.3.1.1. 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% ( $1.01/6 \times 100$ ) from Students' actual Monitoring Accuracy, on a scale ranging from zero to six (before self-scoring, see measure C-SMA-TJ in Table 6.2).

#### 6.3.1.2. Identifying Students with Substantially Inaccurate Monitoring Judgments (RQ1B)

As can be derived from the numbers presented in Table 6.3, 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 performance well (which are those who would be most in need of support).

## 6.3.2. Teacher Judgments of Students' Regulation Accuracy (RQ2)

### 6.3.2.1. 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% 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 6.2).

**After Self-scoring.** On average, Student Regulation Accuracy according to Teacher Judgments deviated 0.48 or 24.00% 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 6.2).

### 6.3.2.2. Identifying Students with Inaccurate Regulation Judgments (RQ2B)

**Before Self-scoring.** As can be derived from Table 6.4, 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 6.5, 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 6.2**

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

Measure	Calculation
Student Performance (SP)	
Student Need for intervention (SN)	
Student Monitoring Judgment (SMJ)	
Student Regulation Judgment (SRJ)	
Student Monitoring Accuracy (SMA)	Absolute difference between Student Monitoring Judgment (SMJ) and Student Performance (SP)
Student Regulation Accuracy (SRA)	Absolute difference between Student Regulation Judgment (SRJ) and Student Need for intervention (SN)
Teacher Judgment of Student Performance (TJSP)	
Teacher Judgment of Student Need for intervention (TJSN)	
Teacher Judgment of Student Monitoring Judgment (TJSMJ)	
Teacher Judgment of Student Regulation Judgment (TJSRJ)	
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).
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)
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)
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)

<sup>a</sup> 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 6.3.3.1.

<sup>b</sup> Values closer to zero indicate more accurate or correct judgments.

	<b>Range</b>	<b><i>n</i></b>	<b><i>M (SD)</i></b>
	0 to 6	764	3.20 (1.95)
	0 to 2	751	Before Self-scoring: 1.20 (0.84) <sup>a</sup>
		746	After Self-Scoring: 1.20 (0.84) <sup>a</sup>
	0 to 6	760	3.79 (1.71)
	0 to 2	755	Before Self-scoring: 0.90 (0.80)
		739	After Self-scoring: 0.91 (0.83)
	0 to 6 <sup>b</sup>	760	1.15 (1.17)
	0 to 2 <sup>b</sup>	744	Before Self-scoring: 0.54 (0.69)
		728	After Self-scoring: 0.39 (0.58)
	0 to 6	764	3.86 (1.70)
	0 to 2	764	0.92 (0.84)
	0 to 6	764	4.02 (1.64)
	0 to 2	764	Before Self-scoring: 0.94 (0.84)
		762	After Self-scoring: 0.95 (0.81)
	0 to 6 <sup>b</sup>	764	0.78 (0.83)
	0 to 2 <sup>b</sup>	764	Before Self-scoring: 0.42 (0.60)
		762	After Self-scoring: 0.31 (0.53)
	0 to 6 <sup>b</sup>	760	1.01 (1.01)
	0 to 2 <sup>b</sup>	744	Before Self-scoring: 0.65 (0.66)
		726	After Self-scoring: 0.48 (0.60)

**Table 6.3**

*Contingency Table of Student Monitoring Accuracy and Student Monitoring Accuracy According to Teacher Judgments*

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

**Table 6.4**

*Contingency Table of Student Regulation Accuracy and Student Regulation Accuracy According to Teacher Judgments Before Self-Scoring*

<b>Student Regulation Accuracy before self-scoring</b>	<b>Student Regulation Accuracy according to Teacher Judgments before self-scoring</b>			<b>Total</b>
	<b>0</b>	<b>1</b>	<b>2</b>	
<b>0</b>	265	131	26	422
<b>1</b>	160	67	12	239
<b>2</b>	53	23	7	83
<b>Total</b>	478	221	45	744

**Table 6.5**

*Contingency Table of Student Regulation Accuracy and Student Regulation Accuracy According to Teacher Judgments After Self-Scoring*

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

### 6.3.3. Relation Between Perceived Student Characteristics and Student Monitoring/Regulation Accuracy According to Teacher Judgments (RQ3)

#### 6.3.3.1. Monitoring

Table 6.6 shows the effects of the teachers' perceptions of student characteristics on Student Monitoring/Regulation Accuracy according to Teacher Judgments. Teachers' perceptions of students' conscientiousness ( $B = -0.21, p = .009$ ), general mathematics ability ( $B = -0.19, p = .004$ ), and nationality ( $B = 0.06, p = .045$ ) had a significant effect on Student Monitoring Accuracy according to Teacher Judgments. The direction of the effects (positive/negative) indicated that the teachers thought that their students made more accurate monitoring judgments when they perceived their students to work more conscientiously, have higher general mathematics abilities, and have 'more of a Western background' (i.e., students and their parents born in Western countries)<sup>4</sup>. The standardized effects of students' conscientiousness ( $\beta = -0.34$ ) and mathematics ability ( $\beta = -0.40$ ) were considerably larger than the standardized effect of students' nationality ( $\beta = 0.10$ ). The effect size of all student characteristics together, in terms of  $f^2$ , was 0.64, indicating a large effect of teachers' perceptions of student characteristics on their 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).

#### 6.3.3.2. Regulation Before Self-scoring

Students' extraversion ( $B = 0.10, p = .008$ ) and students' self-concept ( $B = -0.13, p \leq .001$ ) as perceived by the teachers had a significant effect on Student Regulation Accuracy according to Teacher Judgments before self-scoring (Table 6.6). When teachers' perceptions of students' extraversion *decreased* by one standard deviation, Student Regulation Accuracy according to Teacher Judgments before self-scoring increased (i.e., teachers thought that students made more accurate regulation judgments) by 0.26 standard deviations. When teachers' perceptions of students' self-concept *increased* by one standard deviation, Student Regulation Accuracy according to Teacher Judgments before self-scoring increased by 0.31 standard deviations. The effect size of all student characteristics together, in terms of  $f^2$ , was 0.35, indicating a large effect (Cohen, 1988) of teachers' perceptions of student characteristics on their judgments of how well students regulated their learning prior to self-scoring.

---

4 Teacher-perceived students' nationality was coded as follows (see also Table S6.1 in the Supplementary Materials): (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 6.6**

*The Effects of Teacher-Perceived Student Characteristics on Student Monitoring and Regulation Accuracy According to Teacher Judgments*

	Monitoring N = 593	Regulation	
		Before self-scoring N = 589	After self-scoring N = 591
Fixed effects	B (SE)	B (SE)	B (SE)
Intercept	1.24 (0.31)***	0.21 (0.20)	0.38 (0.12)***
Fixed effects student level			
Amount of contact	0.10 (0.07)	0.11 (0.06)	0.02 (0.05)
Conscientiousness	-0.21 (0.08)**	0.01 (0.04)	0.01 (0.04)
Effort	0.07 (0.07)	-0.01 (0.04)	-0.00 (0.04)
Extraversion	0.02 (0.04)	0.10 (0.04)**	0.05 (0.03)
Interest	0.15 (0.10)	-0.05 (0.07)	-0.06 (0.06)
Learning problems	0.14 (0.09)	0.02 (0.05)	0.08 (0.06)
Likeability	0.01 (0.06)	-0.02 (0.05)	-0.01 (0.05)
Mathematics ability	-0.19 (0.07)**	0.02 (0.04)	0.02 (0.04)
Nationality <sup>a</sup>	0.06 (0.03)*	0.04 (0.03)	0.05 (0.03)
Self-concept	-0.12 (0.08)	-0.13 (0.04)***	-0.06 (0.04)
Sex	0.04 (0.08)	0.09 (0.05)	0.01 (0.05)
Random effects	SS (SE)	SS (SE)	SS (SE)
$\sigma^2_e$ (task)	0.44 (0.06)***	0.24 (0.03)***	0.18 (0.02)***
$\sigma^2_{u0}$ (student)	0.14 (0.06)*	0.08 (0.03)***	0.09 (0.03)***
$\sigma^2_{v0}$ (teacher)	0.03 (0.03)	0.03 (0.02)	0.00 (0.01)
R <sup>2</sup> (student level)	0.39	0.26	0.09

<sup>a</sup> See for coding footnote 4. A higher value means a 'less Western background'.

\*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

### 6.3.3.3. Regulation After Self-scoring

None of the (perceived) student characteristics significantly affected 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, indicating a small to medium effect of teachers' perceptions of student characteristics on their judgments of how well students regulated their learning after self-scoring (Cohen, 1988).

### **6.3.4. Relation Between Perceived Student Characteristics and Correctness of 'Student Monitoring/Regulation Accuracy According to Teacher Judgments' (RQ4)**

Finally, we wanted to explore for which students it is most easy for teachers to make accurate judgments about their monitoring/regulation skills. We therefore analyzed whether and to what extent student characteristics as perceived by the teachers explain the degree of Correctness of 'Student Monitoring/Regulation Accuracy according to Teacher Judgments'. These results are displayed in Table 6.7.

#### **6.3.4.1. Monitoring**

None of the student characteristics significantly affected the Correctness of 'Student Monitoring Accuracy according to Teacher Judgments' (Table 6.7). The effect size of all student characteristics together, in terms of  $f^2$ , was 0.12, indicating a small to medium effect of teachers' perceptions of student characteristics on the correctness of their judgments of how well students monitored their learning (Cohen, 1988).

#### **6.3.4.2. Regulation Before Self-scoring**

Only students' nationality, as perceived by teachers, had a significant and positive effect on the Correctness of 'Student Regulation Accuracy according to Teacher Judgments' before self-scoring ( $B = 0.09$ ,  $p \leq .001$ ; Table 6.7). This effect was positive, meaning that the Correctness of 'Student Monitoring Accuracy according to Teacher Judgments' increased (i.e., the value became closer to zero) when teachers perceived students to have a more Western background (see footnote 4 for how we coded students' nationality;  $\beta = 0.29$ ). The effect size of all student characteristics together, in terms of  $f^2$ , was 0.22, indicating a medium effect of teachers' perceptions of student characteristics on the correctness of their judgments of how well students regulated their learning prior to self-scoring (Cohen, 1988).

To gain more insight into the relationship between students' nationality and teacher judgments of student regulation skills before self-scoring, we additionally explored the means of students' nationality per combination of Students' Regulation Accuracy and Student Regulation Accuracy according to Teacher Judgments, see Table 6.8. As Table 6.8 shows, the students for whom the Student Regulation Accuracy was in line with the Student Regulation Accuracy according to Teacher Judgments (Correctness of 'Student Regulation Accuracy according to Teacher Judgments' = 0) relatively often had more of a Western background. Students who made accurate regulation judgments before self-scoring (Student Regulation Accuracy = 0) of whom the teachers thought that these students were not correct (Student Regulation Accuracy according to Teacher Judgments = 1 or 2) had relatively often a non-Western background.

**Table 6.7**

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

	Monitoring N = 589	Regulation	
		Before Self-scoring N = 576	After Self-scoring N = 556
Fixed effects	B (SE)	B (SE)	B (SE)
Intercept	1.00 (0.38)**	0.76 (0.28)**	0.49 (0.24)*
Fixed effects student level			
Amount of contact	0.10 (0.06)	0.04 (0.07)	0.08 (0.07)
Conscientiousness	-0.03 (0.07)	-0.06 (0.05)	-0.08 (0.05)
Effort	-0.08 (0.11)	0.02 (0.07)	-0.01 (0.08)
Extraversion	-0.02 (0.05)	0.03 (0.04)	0.02 (0.04)
Interest	0.02 (0.07)	-0.05 (0.07)	-0.05 (0.07)
Learning problems	-0.14 (0.12)	-0.04 (0.08)	-0.05 (0.07)
Likeability	0.03 (0.06)	-0.03 (0.05)	-0.04 (0.04)
Mathematics ability	-0.03 (0.05)	0.03 (0.03)	0.01 (0.04)
Nationality <sup>a</sup>	0.03 (0.07)	0.09 (0.02)***	0.01 (0.03)
Self-concept	0.05 (0.07)	-0.03 (0.05)	0.03 (0.05)
Sex	-0.02 (0.11)	0.04 (0.06)	0.09 (0.07)
Random effects	SS (SE)	SS (SE)	SS (SE)
$\sigma_e^2$ (task)	0.89 (0.12)***	0.39 (0.03)***	0.31 (0.04)***
$\sigma_{u0}^2$ (student)	0.09 (0.10)	0.05 (0.03)	0.04 (0.03)
$\sigma_{v0}^2$ (teacher)	0.00 (0.02)	0.00 (0.01)	0.00 (0.01)
R <sup>2</sup> (student level)	0.11	0.18	0.16

<sup>a</sup> See for coding footnote 4. A higher value means a 'less Western background'.

\*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

### 6.3.4.3. Regulation After Self-scoring

None of the student characteristics as perceived by teachers significantly affected the Correctness of 'Student Regulation Accuracy according to Teacher Judgments' after self-scoring (Table 6.7). The effect size of all student characteristics together, in terms of  $f^2$ , was 0.19, indicating a medium effect of teachers' perceptions of student characteristics on the correctness of their judgments of how well students regulated their learning after self-scoring (Cohen, 1988).

**Table 6.8**

*Means and Standard Deviations (Between Brackets) of Students' Nationality per Combination of Student Regulation Accuracy and Student Regulation Accuracy According to Teacher Judgments, Before Self-scoring*

	<b>Student Regulation Accuracy according to Teacher Judgments before self-scoring</b>		
	<b>0</b>	<b>1</b>	<b>2</b>
<b>Student Regulation Accuracy before self-scoring</b>			
<b>0</b>	<i>n</i> = 231 0.19 (0.68)	<i>n</i> = 119 0.32 (0.90)	<i>n</i> = 23 0.43 (1.20)
<b>1</b>	<i>n</i> = 136 0.25 (0.80)	<i>n</i> = 60 0.15 (0.63)	<i>n</i> = 12 0.00 (0.00)
<b>2</b>	<i>n</i> = 45 0.24 (0.74)	<i>n</i> = 20 0.55 (1.28)	<i>n</i> = 7 0.14 (0.38)

*Note.* The mean of students' nationality in the total sample is 0.24 (0.78).

## 6.4. 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 affected by (perceived) student-related factors.

### 6.4.1. 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 (2022) 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 6.2, 6.4, and 6.5). 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.

#### **6.4.2. 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 might use their perceptions of students' mathematics abilities, variables related to students' working behavior (such as effort and conscientiousness), and students' self-concept when making judgments of students' monitoring and regulation skills. Some of these expected relations indeed appeared. The teachers seemed to think that: (1) students who work more conscientiously and are more skilled in mathematics would make more accurate monitoring judgments, and (2) students who have a higher self-concept would make more accurate regulation judgments prior to self-scoring. Unexpectedly, the teachers also seemed to think that students with a 'more Western background' would make more accurate monitoring judgments and that students who were less extravert would make more accurate regulation judgments prior to self-scoring. We should note that all these unique effects 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 large amount

of variance in teacher judgments of students' monitoring and regulation accuracy before self-scoring. This suggests that teachers' overall picture of their students might influence their judgments of students' monitoring and regulation skills before self-scoring. 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 6.5), and (2) the effect of student characteristics on teacher judgments of students' regulation accuracy after self-scoring decreased to a small to medium effect size (compared to a large effect size before self-scoring). In other words, teachers seemed to think that students' characteristics would 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 had a unique significant effect on the accuracy of teachers' judgments of students' monitoring and regulation skills, except for students' nationality. Additional explorations pointed out that students whose regulation skills prior to self-scoring were maximally underestimated by their teachers were relatively often from non-Western origin. We found no support for the hypothesis that teachers made more accurate judgments about students for whom they have *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 effects of (perceived) student characteristics considered individually were not significant, except for nationality, our findings indicate that the accuracy of teacher judgments of students' monitoring and regulation skills might be somewhat influenced by their perceptions of student characteristics: Considered together, the characteristics explained a small to medium amount of variance in the accuracy of teacher judgments of students' monitoring skills, and medium amounts of variance in the accuracy of teacher judgments of students' regulation skills before and after self-scoring. Possibly, it might be easier for teachers to make accurate judgments for students with a specific profile of characteristics, compared to other students, and future research should illuminate what these specific profiles look like.

### 6.4.3. 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.

As for the information 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.

Future research should also establish to what extent our findings would 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.

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 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.

#### **6.4.4. Conclusions and Implications**

Primary school students often make inaccurate monitoring and regulation judgments (Baars et al., 2014a; Oudman et al., 2022b; Prinz et al., 2020; Van Loon & Roebbers, 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., 2022b; Van Loon & Roebbers, 2017). Hence, it 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. This implies that we would need to find ways to help teachers identify those students. Moreover, our findings suggest that teachers' judgments of students' regulation accuracy might be somewhat biased by students' nationality, and it would be important to find ways to counteract such biases





# Chapter 7

## Summary and Discussion

The main aim of the research presented in this dissertation was to gain more insight into how the accuracy of primary school students' and teachers' monitoring and regulation judgments of students' performance on mathematical problem-solving tasks can be improved. Monitoring judgments concern evaluations of individual students' performance. Regulation judgments refer to decisions about what subsequent activities (e.g., restudy, additional practice, or additional instruction) students should engage in to improve their performance. As such, making accurate monitoring judgments is a necessary (though not always sufficient) condition for making accurate regulation judgments (e.g., Dunlosky & Rawson, 2012; Van de Pol et al., 2011).

Making accurate monitoring judgments is necessary for *teachers* to make instructional decisions that are adapted to the diverse needs of individual students, also referred to as 'adaptive' or 'differentiated' instruction (Parsons et al., 2018; Tomlinson et al., 2003; Van de Pol et al., 2011). At the same time, *students* have to learn to accurately monitor and regulate their learning, as this is a pre-requisite for effective self-regulated learning and the development of these skills lays the foundation for students' lifelong learning (Bjork et al., 2013; OECD, 2022) and beneficially influences their academic achievement (Dent & Koenka, 2016). Consequently, teachers are also increasingly expected to help students develop their monitoring and regulation skills (Dignath & Büttner, 2018).

Prior research has shown, however, that primary school students' monitoring and regulation judgments are often inaccurate, for instance when memorizing information (e.g., Roebers et al., 2014), when learning from texts (e.g., De Bruin et al., 2011), and when learning to solve problems (e.g., Baars et al., 2014a; Boekaerts & Rozendaal, 2010; García et al., 2016). Moreover, teachers' monitoring judgments of students' performance are often inaccurate, for instance, research has shown room for improvement in the accuracy of teacher judgments of their students' vocabulary, reading comprehension, and mathematical problem-solving skills (for review studies, see Kaufman, 2020; Südkamp et al., 2012; Urhahne & Wijnia, 2021). Because accurate monitoring is a necessary condition for accurate regulation (for teachers: Van de Pol et al., 2011; for students: Dunlosky & Rawson, 2012), this is problematic, because it means that teachers' instructional decisions and students' self-regulated learning behavior are not always optimally adapted to students' needs.

An explanation for the inaccuracy of students' and teachers' monitoring judgments lies in the information or *cues* on which the judgments are based. Students (Koriat, 1997) and teachers (Byers & Evans, 1980; Snow, 1968) use a variety of cues when making judgments about students' current level of performance or degree of understanding. The more predictive or *diagnostic* of students' actual performance the cues being used, the more accurate a student's or teacher's monitoring judgment will be (Koriat, 1997; Thiede et al., 2010, 2015; Van Loon et al., 2014). The reason why monitoring judgments are often inaccurate, is that the available cues differ in the extent to which they are

diagnostic of students' actual performance, and students and teachers do not always use highly diagnostic cues. Therefore, if we ultimately want to improve self-regulated learning and adaptive instruction, it is important to investigate whether interventions that give students and teachers access to more diagnostic cues, improve their monitoring accuracy, and, in turn, their regulation accuracy.

The aim of the studies presented in this dissertation was to investigate how primary school students' and teachers' judgments of students' performance on mathematical problem-solving tasks were affected by interventions that focused their attention on cues that are more diagnostic of students' actual performance. More specifically, it was investigated whether and how self-scoring their problem solutions improved students' (awareness of) their monitoring and regulation accuracy, whether and how providing teachers with information on students' prior practice task performance improved their (awareness of their) monitoring judgments of students' performance, and how well teachers were able to make correct estimations of students' monitoring and regulation accuracy.

This dissertation contains five empirical studies, presented in Chapters 2 to 6, that were conducted in Dutch grade 6, with 9- to 10-year-old students and their teachers. The data for chapters 2, 3, 5, and 6 were gathered in one large data collection; chapter 4 concerned a separate data collection. In this final chapter, I will first summarize and then discuss the main findings from the studies reported in this dissertation. Subsequently, I will discuss limitations of the research presented in this dissertation and important directions for future research. Finally, I will present (potential) implications for educational practice, a theme that is close to my heart since I am also a primary school teacher and conducted this dissertation research as a 'PromoDoc', that is, working part-time at the university as a PhD candidate and part-time as a teacher in primary education.

## 7.1. Summary of the Main Findings

The studies presented in **Chapter 2 and 3** focused on improving students' (awareness of their) monitoring and regulation accuracy. The study described in **Chapter 2** investigated whether self-scoring would improve primary school students' monitoring and regulation accuracy when solving procedural mathematics problems; more specifically, multiplication and division problems. Self-scoring entails comparing one's own answers to a standard of the correct answers, which gives students access to highly diagnostic *performance cues* (Rawson & Dunlosky, 2007). Furthermore, it was investigated whether the effects of self-scoring on (potential) improvements in monitoring and regulation accuracy would differ between low- and high-performing students, as this would call for a differential focus in interventions. Students' absolute monitoring accuracy was measured by computing the absolute difference (i.e., regardless of whether it was positive or negative) between

students' judgments of how many problems they answered correctly and the number of problems they actually answered correctly. Students' absolute regulation accuracy was measured by computing the absolute discrepancy between students' own judgments of their need for additional instruction or practice and their actual need (as judged by experts) for additional instruction or practice on the multiplication or division problems.

The results showed that having primary school students self-score their solutions to procedural mathematics problems is an effective way to increase their monitoring accuracy: Most students were fully accurate after self-scoring. Importantly, this also translated into more accurate regulation judgments (in the overall sample on both tasks and in the low-and high-performing subsamples on the multiplication task). As for monitoring, high-performing students made more accurate judgments than low-performing students before self-scoring, but after self-scoring this difference disappeared on the division task and became very small—although it remained significant—on the multiplication task. As for regulation, high-performing students made substantially more accurate judgments than low-performing students, both before and after self-scoring on the multiplication task (but not on the division task). Thus, self-scoring seems to be an effective intervention to improve monitoring accuracy as well as regulation accuracy, although many students—and especially low-performing students—may still need additional support for making accurate regulation judgments.

However, only improving students' monitoring and regulation accuracy may not be sufficient to impact their actual self-regulated learning behavior. Students' *awareness* of their judgment (in)accuracy might also be an important predictor of how they will self-regulate their learning, because accuracy awareness might affect whether and how they act upon their judgments (as suggested by Gabriele et al., 2016; Händel & Fritzsche, 2016; Patterson et al., 2001). Therefore, the study presented in **Chapter 3** explored whether students were aware of their monitoring and regulation (in)accuracy and whether and how this was affected by self-scoring.

After making a monitoring or regulation judgment, students had been asked to rate their feeling of confidence in the accuracy of their monitoring (before self-scoring) and regulation (before and after self-scoring) judgments. During a pilot study, students also rated their confidence in their monitoring accuracy after self scoring. Because all students felt maximally confident and experienced the question as "strange", this question was removed in the main study.

Students show accuracy awareness when they feel relatively more confident about the accuracy of more accurate judgments than of less accurate judgments and vice versa. The findings indicated that on average, students showed limited awareness of their monitoring and regulation (in)accuracy prior to self-scoring. Overall (i.e., in the entire sample), self-scoring seemed to improve students' regulation accuracy awareness. Yet it had limited effect on regulation accuracy awareness for low-performing students and

for students whose regulation accuracy decreased or stayed equally inaccurate after self-scoring. Thus, future research on additional or other means to increase students' regulation accuracy awareness (e.g., feedback or training) is needed.

The studies reported in **Chapter 4, 5, and 6** focused on improving teachers' (awareness of their) monitoring judgments of students' performance, as well as on their judgments of students' monitoring and regulation accuracy. In the studies reported in **Chapter 4 and 5** the availability of performance cues (i.e., students' performance on a prior practice task) and student cues (i.e., general student characteristics such as their interest in mathematics or nationality) was manipulated. It was expected that providing primary school teachers with student products from which they could infer more diagnostic *performance cues* would affect their cue use and improve the accuracy of their monitoring judgments of students' performance.

The study presented in **Chapter 4** was concerned with teacher judgments of their students' conceptual understanding of decimal magnitude. The students completed a practice task and follow-up task, while teachers only made judgments about the students' performances on the follow-up task. All teachers made these judgments under three conditions, while having access to: (1) only students' names (i.e., only student cues available), (2) only anonymized students' answers on decimal magnitude practice problems (i.e., only performance cues available), and (3) both students' names and their answers (i.e., both student and performance cues available). Knowing the students' names would give teachers access to student cues based on their knowledge of the student. These student cues were expected to have low diagnosticity for students' performance on a specific task, in this case: decimal magnitude problems. Having the students' answers on the practice problems available, would allow teachers to infer students' decimal (mis) conceptions—cues that were expected to have high diagnosticity—by analyzing error patterns in students' answers. Teachers made item-specific judgments of students' conceptual understanding of decimal magnitude. The accuracy of these judgments was measured in terms of sensitivity (correctly judging what students did understand) and specificity (correctly judging what students did not understand). The teachers were asked to think aloud while making judgments, to measure their cue use.

The findings indicated that giving teachers access to the students' answers in addition to students' names, did not significantly improve the sensitivity or specificity of teacher judgments of students' decimal magnitude understanding. This can possibly be explained by the finding that the teachers—despite the availability of performance cues—did not use less student cues than in the name only condition, as shown by the think-aloud data. Giving teachers access to students' answers only (i.e., without students' names, so they could not rely on student cues), did improve the specificity, but not the sensitivity of teacher judgments.

Identifying the decimal (mis)conceptions from students' answers in the study presented in Chapter 4 would require quite some interpretation by the teachers. Having

products available that do not ask for complex interpretations, might make it easier for teachers to use diagnostic performance cues and *ignore* non-diagnostic student cues, which should improve the accuracy of their monitoring judgments.

The study described in **Chapter 5** tested this hypothesis, by providing teachers with performance cues that do not ask for interpretation: students' scores on prior tasks. All teachers made judgments of how their students would perform on a multiplication and division task under two conditions: while having access to: (1) only students' names (i.e., only student cues available), and (2) both students' names and their scores on similar multiplication and division tasks, completed one week earlier (i.e., student and performance cues available). Teachers' absolute monitoring accuracy was measured by computing the absolute discrepancy between a teacher's prediction of how many problems a student answered correctly on the multiplication or division task and student's actual performance on that task. An indication of teachers' cue use was obtained by computing correlations between teacher judgments of their students' performance and measures of teachers' perceptions—measured by a questionnaire—of general student characteristics (e.g., interest in mathematics, nationality).

It also goes for teachers that, besides making accurate monitoring judgments, it is important that they are *aware* of their judgment (in)accuracy, as this might make their instructional decisions more effective (Gabriele et al., 2016). Therefore, it was also explored to what extent teachers would be aware of their monitoring (in)accuracy and whether and how this was affected by the availability of performance cues. Teachers' accuracy awareness was measured by asking them to rate their feeling of confidence in the accuracy of their monitoring judgments.

The findings showed that giving teachers access to students' scores on prior tasks in addition to student names positively affected teachers' absolute monitoring accuracy of their students' performance on procedural mathematics tasks. The correlational cue-use analyses suggest that this effect came about because teachers not only used the performance cues but also seemed to ignore the less diagnostic student cues when the prior task scores were also available. Furthermore, the findings showed that when only having access to student cues, teachers were somewhat aware of their (in)accuracy, in that they feel relatively more confident about more accurate judgments and relatively less confident about less accurate judgments. Teachers' confidence in and awareness of their monitoring (in)accuracy increased further when they were given access to students' prior task scores.

The study presented in **Chapter 6** explored how well teachers would be able to judge the accuracy of their students' monitoring and regulation skills. Making accurate judgments of students' monitoring and regulation skills is needed to be able to identify those students who make substantially inaccurate judgments and who would therefore be most in need of support with developing their monitoring and regulation skills. For this study, the teachers were not only asked to judge their students' performance and needs

(i.e., additional practice or additional instruction), but were also asked to indicate what monitoring and regulation judgments they thought their students had made. From these judgments, variables were computed that indicated (the accuracy of) teacher judgments of students' monitoring and regulation accuracy. Moreover, it was explored whether and how (the accuracy of) teacher judgments of students' monitoring and regulation accuracy would be affected by teachers' perceptions of general student characteristics, to gain more insight into how (in)accurate teacher judgments of student monitoring and regulation skills are established.

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 judgments, while for those students 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, suggesting that teachers' overall picture of their students might influence (the accuracy of) these judgments. Moreover, teacher judgments of students' regulation accuracy before self-scoring seemed somewhat biased by students' nationality: The teacher judgments were on average more accurate if the students had a more Western Background. Thus, these findings show that teachers might need help with identifying students who make substantially inaccurate monitoring and regulation judgments, and with counteracting potential biases in their judgments, to be able to optimally support their students' development of monitoring and regulation skills.

## **7.2. Discussion of Main Findings**

The studies reported in Chapters 2 to 6 provide new insights into how the accuracy of primary school students' monitoring and regulation judgments and teachers' monitoring judgments of students' performance on problem-solving tasks can be improved. Below, I will discuss the main findings in the context of the extant literature.

### **7.2.1. Improving Students' Monitoring and Regulation Accuracy**

Prior research had shown that self-scoring one's work with the use of standards, which is common practice in many Dutch primary schools, improved primary school students' monitoring and regulation accuracy when learning concepts (Van Loon & Roebbers, 2017) and secondary school students' monitoring accuracy when solving biology problems (Baars et al., 2014b). The study presented in Chapter 2 shows that self-scoring is also an effective way to increase primary school students' monitoring accuracy when solving procedural mathematics problems (most students' monitoring judgments became fully accurate), and, importantly, that this also leads to improved accuracy of regulation judgments.



Nevertheless, there was still room for further improvement in regulation accuracy even after self-scoring; indicating that many students—and especially low-performing students—still need help from their teachers with making accurate regulation judgments.

Unfortunately, the findings from Chapter 6 show that teachers had difficulties identifying those students who made substantially inaccurate monitoring and regulation judgments, before and after self-scoring, while those students would be most in need of their teacher's support. To the best of my knowledge, this study was the first to investigate whether teachers can tell how well students can monitor and regulate their own learning process: In another recent study, we also looked at teacher judgments of students monitoring accuracy—not regulation accuracy—on text comprehension tasks in secondary education (Van de Pol & Oudman, 2022). The findings reported in Chapter 6 indicate that teachers are quite adept at recognizing which students can monitor their performance well (which was also the case in the study by Van de Pol & Oudman), but that they need help with identifying students who need support in the development of self-regulated learning skills, for instance by means of learning analytics displayed in teacher dashboards (see section 7.4.3).

### **7.2.2. Improving Teachers' Monitoring Accuracy**

With regard to teachers' ability to monitor their students' performance (a prerequisite for adaptive instruction; e.g., Klug et al., 2013; Van de Pol et al., 2011), research has shown substantial room for improvement in teachers' monitoring judgment accuracy (Kaufman, 2020; Südkamp et al., 2012; Urhahne & Wijnia, 2021). Yet, research that provides insight into how teachers' monitoring accuracy can be improved was scarce when I started this dissertation research. As for students, the key to improving monitoring accuracy would seem to lie in providing more diagnostic performance cues. A study by Thiede et al. (2015) provided some evidence that focusing more on performance cues would improve the accuracy of teacher judgments of students' mathematical performance, but their intervention was part of a larger professional development program and they did not measure what cues teachers used. More recent studies on formative assessment practices, aimed at collecting performance cues to guide instructional decisions, showed mixed evidence regarding improvements in the accuracy of primary school teachers' monitoring judgments of their students' mathematical performance (Thiede, 2018, 2019; Zhu & Urhahne, 2018).

The findings from Chapters 4 and 5 suggest that explanations for these mixed findings might lie in how easily teachers can infer—and thus use—performance cues from student products, and in whether they can simultaneously ignore less diagnostic student cues (which can be difficult as those are always available in educational practice). The study reported in Chapter 4 showed that when the available more diagnostic performance cues ask for complex interpretations, teachers continued to rely on student cues as

much as when they had no performance cues available. Similar results were found by Van de Pol et al. (2021), in a study on teachers' monitoring of secondary school students' text comprehension. The findings from the study described in Chapter 5 suggest that when performance cues are easy to infer (e.g., from students' scores on similar prior tasks) teachers do seem to be able to not only use the more diagnostic performance cues but also ignore the less diagnostic student cues. Future research should verify our assumptions regarding the effects of ease of inferring performance cues, which would require direct comparisons within a study. Moreover, future research should identify what types of products or interventions would provide teachers with easy-to-use performance cues or could help them to learn to ignore less diagnostic student cues.

### **7.2.3. Students' and Teachers' Awareness of Their Judgment (In)Accuracy**

Much of the (recent) research in educational sciences has focused on finding interventions to improve primary school students' monitoring and regulation accuracy (e.g., Baars et al., 2014a; De Bruin et al., 2011; Kostons & De Koning, 2017; Van Loon & Roebbers, 2017) and primary school teachers' monitoring accuracy (Thiede et al., 2015, 2018, 2019; Zhu & Urhbane, 2018). However, for acting upon their judgments (and thus, for the effectiveness of teaching and learning activities), it has been argued that it is important that students and teachers are aware of the (in)accuracy of their judgments: They will presumably only act upon accurate judgments when they are aware of (i.e., feel confident about) the accuracy of these judgments (as suggested by Gabriele et al., 2016; Händel & Fritzsche, 2016; Patterson et al., 2001; Praetorius et al., 2013). Yet, whether and to what extent primary school students and teachers are aware of their judgment (in)accuracy, and whether and how their accuracy awareness could be increased, had not yet been addressed: The little prior research available focused only on adolescent and adult students, and only on the status quo, not on improving it (Fritzsche et al., 2018; Händel & Dresel, 2018; Nederhand et al., 2021).

The results reported in Chapters 3 and 5 give some first answers to these questions. The findings indicate that, prior to intervention, primary school teachers did show awareness of their monitoring accuracy (teachers' regulation accuracy awareness was not measured), but their students did not show awareness of their monitoring and regulation accuracy. This raises the question of when (around what age) students start to develop accuracy awareness. That is, prior research has shown that adult students did show accuracy awareness (Fritzsche et al., 2018; Händel & Dresel, 2018), but suggested that adolescents did not (Nederhand et al., 2021), although the research with adolescents did not use repeated measures—which is needed to draw conclusions about students' accuracy awareness (see Chapter 3). Interestingly and importantly, the findings from Chapters 3 and 5 also showed that interventions that focused students' and teachers'

attention on diagnostic performance cues (by asking students to self-score their work or provide teachers with students' prior task scores) increased their confidence in and their awareness of their judgment (in)accuracy. This is good news, as improved accuracy awareness presumably increases the chance that students and teachers actually act upon accurate judgments, which might ultimately lead to more learning progress, although this should be confirmed by future research (see the following section 7.3)

### **7.3. Limitations and Suggestions for Future Research**

As study specific limitations have been addressed in the respective chapters, I will focus on overarching issues here, which concern the measurement of regulation, effects of the type of the to-be-judged task, the measurement of cue use and diagnosticity, and, ultimately, testing the effectiveness of the interventions investigated in this dissertation for adaptive teaching and self-regulated learning behavior.

The research presented in this dissertation used measures of regulation judgments on problem-solving tasks that are highly relevant for teaching and learning in primary school: In the studies in Chapter 2, 3, and 6, students and teachers were asked to indicate whether they thought that they or their students would need additional practice or additional instruction on the multiplication or division problems. However, in primary schools, students can also perform other learning activities when they feel they did not yet master a learning goal, such as asking a peer for help or studying a worked example. Future research might therefore consider elaborating on our way of measuring regulation judgments on problem-solving tasks. It is an open question whether this would also explain why there was still room for improvement in students' regulation accuracy after the self-scoring intervention—even though monitoring accuracy was almost perfect—in the study presented in Chapter 2. As prior studies on self-scoring had similar findings regarding regulation accuracy (e.g., Baars et al., 2014b; Van Loon & Roebbers, 2017), it would be interesting for future research to attempt to acquire more insight into students' motives for regulation decisions after self-scoring, by using think aloud protocols or interviews. This might lead to ideas of how to help students to regulate their learning more effectively, in line with their highly accurate monitoring judgments after self-scoring.

In educational practice, as well as in the different studies presented in this dissertation, students and teachers make judgments at different levels of granularity and on different types of tasks. The studies reported in Chapter 2, 3, 5, and 6 were concerned with whole-task judgments of procedural problem-solving performance (multiplication and division), while the study presented in Chapter 4 was concerned with item-specific judgments of conceptual understanding (of decimal magnitude). As the findings from Chapters 2 and 3 show, and previous studies have also shown (Boekaerts & Rozendaal, 2010;

Rutherford, 2017), students' monitoring and regulation accuracy (awareness) seems to vary across different types of mathematical tasks. Similarly, teachers' monitoring accuracy has been shown to vary across different content domains within mathematics (Kolovou et al., 2021; Thiede et al., 2015). One possible explanation for such differences might be that for some tasks knowledge gaps are easier to identify (see the discussion section of Chapter 2), but future research should attempt to gain more insight into why these differences arise. Moreover, grain size of judgments might influence the (awareness of) monitoring and regulation accuracy: For example, Karst et al. (2018) showed that teacher monitoring judgments were more accurate on the item level than on the whole-task level. Thus, future research should shed more light on the question of how judgment grain size and task type influence students' and teachers' cue use, monitoring and regulation accuracy, and their accuracy awareness.

Another important issue is how to measure students' and teachers' cue use. Most prior studies, as well as the studies reported in Chapter 5 and 6, inferred students' or teachers' cue use from correlations between their judgments and the cue values (obtained from student information reported by the participants, or performance information derived from student work). A drawback of this indirect approach is that one cannot draw causal conclusions, because one cannot be sure that the cues that correlate to the judgments were actually used by participants: For instance, they may also have used slightly different cues that are highly related, both conceptually and correlationally, to those that were measured. Moreover, one cannot exclude the possibility that participants might have used entirely different cues that were not measured. Therefore, although it is more time consuming, future research could gain more insight into (effects of interventions on) cue use by moving towards more direct measures of cue use such as think aloud protocols (as used in the study presented in Chapter 4) or short questionnaires that directly ask teachers and students what their judgments were based on (e.g., Van de Pol et al., 2021).

An interesting and relevant question regarding cue use and diagnosticity more generally, is whether these two constructs depend on whether the cues manifest to a specific degree, in a specific direction, or in combination with other cues. For instance, in theory, cues like effort or interest might be diagnostic when they are high or low in a student, but less diagnostic when they are medium/moderate. Or teachers might think that especially students who are good in mathematics *and* invest the effort would be skilled in making accurate regulation judgments. Future research could consider or develop other types of analyses (e.g., latent class analyses) to define different profiles with regard to cue diagnosticity, cue use, and (the accuracy of) students' or teachers' judgments.

The studies presented in this dissertation focused on necessary first steps on the way towards the more distal goal of improving students' self-regulated learning and adaptive teaching in the context of primary school mathematics. It is important in future research to investigate the whole chain, and establish whether (1) helping teachers to focus on

(easy-to-infer) performance cues not only improves their monitoring accuracy, but also leads to more adaptive instructional decisions (i.e., teachers' regulation accuracy), and higher student learning outcomes, and (2) whether self-scoring does not only lead to improvements in students' monitoring and regulation judgment accuracy but also to more effective self-regulated learning behavior in class, as indicated by higher learning outcomes.

In doing so, students' and teachers' accuracy awareness (Chapters 3 and 5, respectively) also needs attention, as the potential importance of accuracy awareness for subsequent decisions and actions is entirely based on theoretical assumptions. Future research should confirm whether or not students and teachers act differently upon their (in)accurate judgments as a function of their confidence in their judgment (in)accuracy. If so, it would be highly relevant to keep looking for interventions that would not only increase students' and teachers' judgment accuracy, but also their awareness of their (in)accuracy.

## **7.4. Implications for Educational Practice and Practice-Oriented Future Research**

### **7.4.1. Implications for Improving Students' Self-Regulated Learning Behavior**

Developing students' self-regulated learning skills is an important objective in many (Dutch) primary schools. However, whereas (in the Netherlands) teacher professional development programs based on scientific insights are widely available for subject areas such as mathematics or "social competence and citizenship", this is not yet the case for fostering students' self-regulated learning. Thus, it is not surprising that both observational studies (Dignath et al., 2018; Kistner et al., 2010; Spruce & Boll, 2015) and self-report studies (De Smul et al. 2019; Van de Velde et al., 2012) have shown that primary school teachers vary widely in how and to what extent they pay attention to self-regulated learning skills in class. Many teachers rarely provide their students with explicit instruction of self-regulated learning skills (for a review, see Dignath & Veenman, 2021).

Yet, there are existing practices that provide excellent opportunities to connect such explicit instructions to. For instance, having students self-score their performance on mathematical problem-solving tasks as well as other tasks, is a common practice in many Dutch primary schools. The findings from the study presented in Chapter 2 clearly show that—at least for procedural mathematics tasks—this seems to be good practice, as it improves students' monitoring accuracy, which is a necessary condition for effective regulation. Although overall, regulation accuracy also improved to some extent after self-scoring, the substantially improved monitoring accuracy did not necessarily result in improved regulation decisions for all students: A finding teachers should be made aware of. Moreover, this finding implies that teachers should more explicitly or more extensively teach their students how to choose for additional practice or instruction, adapted to their

monitoring judgments after self-scoring. In doing so, teachers should take into account that low-performing students might need more support or a different kind of support than high-performing students (cf. Chapter 2). Moreover, if future research would indeed confirm a causal link between students' accuracy awareness and whether or not they act upon their regulation judgments, then teachers should also try to help students become more *aware* of their regulation (in)accuracy after self-scoring, because otherwise, improving students' regulation accuracy might not actually lead to more effective self-regulated learning behavior. Especially low-performing students and students whose regulation decisions are (still) inaccurate after self-scoring would need to become more aware of their (in)accuracy (cf. Chapter 3).

However, "teachers should help students to increase their regulation accuracy (awareness)" is easier said than done: Future practice-oriented research should further investigate how we can provide teachers with specific advice or tools for improving students' regulation accuracy (and accuracy awareness) in different subject areas. For instance, it could be tested whether it is effective to give modeling examples on how to decide whether additional instruction or practice is needed (cf. Kostons et al., 2012; Raaijmakers et al., 2018, who investigated this amongst secondary school students), or by giving students explicit feedback on their regulation accuracy. Possibly, online learning systems could also play a role in this (see section 7.4.3). Then, these findings should be translated, along with findings from prior research on supporting self-regulated learning more generally, into teacher trainings and teacher professional development programs.

### **7.4.2. Implications for Improving Teachers' Instructional Decisions**

The findings from this dissertation also have potential implications for improving teachers' differentiation practices. All too often, teachers use fixed ability grouping—that is, distributing the students in their class into a low-, average-, and high-performing group, based on assessments of their general mathematical abilities completed a few times a year. These groups then get instruction and practice tasks adapted to their level. A problem with such fixed grouping, however, is that students can perform well on one task and poor on another task within the same subject (Nuutila et al., 2018). This may explain why making instructional decisions based on fixed ability groups is not beneficial for primary school students' mathematics performance, especially not for low-performing students (for a meta-analysis, see Deunk et al., 2018). In contrast, flexible grouping, based on students' prior performance, gathered on a daily basis by formative assessment practices, has been shown to be more effective for student learning (Deunk et al., 2018). Note that formative assessments do not have to be presented to students as tests; these can be short activities integrated in the instruction or evaluation phase of each lesson (Wiliam, 2011).

In other words, it is important that teachers base their instructional decisions on students' performance on specific tasks and not on their general abilities. That flexible grouping is not implemented on a large scale might be explained by the finding that teachers often experience formative assessment practices as time consuming and ineffective (for a review, see Yan et al., 2021). That is, substantial interpretation may be needed for the outcomes of formative assessment to be effective for informing their instructional decisions which teachers might not always be able or be inclined to make: This is also shown by the findings reported in Chapter 4, where providing teachers with "complex" performance cues did not increase their monitoring accuracy when student cues were also available. Hence, finding ways to provide teachers with easy to elicit and easy to infer performance cues might be helpful in shifting from fixed to flexible grouping. It would be helpful if mathematics teaching methods would provide teachers with short formative assessment practices and tools to create an overview of students' current level of performance per learning objective. Such tools could consist of tables in which teachers can easily indicate for each student to what extent they master the objective: Some teachers developed such tools themselves, but it would be far more efficient when publishers of the methods would provide such tools. When using online learning systems, such overviews, already filled with data based on students' performance on online practice tasks, can be requested (see the following section 7.4.3).

In addition to helping teachers to focus on diagnostic performance cues, their judgment accuracy (awareness) might also profit from interventions aimed at reducing their focus on non-diagnostic (student) cues, as well as their biases with regard to student characteristics. This could be done, for instance, by providing teachers with knowledge on the diagnosticity of different cues (cf. Rowan et al., 2023), feedback on their own biases with regard to student characteristics (cf. Hadjidemetriou, 2021), or accountability priming (i.e., increasing teachers' responsibility for their decision, cf. Pit-ten Cate et al., 2016).

### **7.4.3. Implications for the Development and Design of Online Mathematics Learning Systems**

Having to elicit diagnostic performance cues for purposes of adaptive instruction, and teaching students how they can regulate their own learning, requires a lot from teachers, who are already wrestling with a very high workload (Gemmink et al., 2020). They can potentially be assisted by online mathematics learning systems, which are increasingly used in primary schools. For instance, in The Netherlands, all of the most widely used primary school mathematics methods nowadays also offer some kind of online platform, and there are dedicated online (mathematics) programs like Rekeningtuin (<https://rekeningtuin.nl>; in English: Math Garden), Snappet (<https://snappet.nl>), or Gynzy (<https://gynzy.com>).

First, such online systems can become promising tools for helping to improve students' monitoring and regulation accuracy (awareness). Many of those systems

make the need for self-monitoring and self-regulation by the student obsolete, by automatically evaluating the students' performance and presenting a new task adapted to their level of performance. Fortunately, researchers, have also started to investigate how online learning systems can be helpful in teaching (primary school) students to make more accurate monitoring and regulation judgments themselves, for instance by giving feedback on students' monitoring and regulation accuracy (e.g., Roll et al., 2011; Molenaar et al., 2020). Next to helping students to improve their monitoring and regulation accuracy, the feedback provided by these systems might also improve their accuracy awareness, although this remains to be investigated.

Second, such online learning systems can also provide information to help teachers to make more adaptive instructional decisions, or even take over such decisions. The systems that are currently used in (Dutch) primary education typically come with a teacher dashboard, that displays performance cues (students' learning analytics data), either in the form of students' item-specific performance, or as students' progress towards mastering a specific learning objective (e.g., addition and subtraction of decimal numbers). However, it might be difficult for teachers to infer diagnostic performance cues from item-specific performance information (cf. Chapter 4) and progress towards a learning objective does not give teachers direct information on how to best help a student. To foster adaptive instructional decisions, it might be most effective if the teacher dashboards could provide aggregated and easy-to-interpret information about what a student does and does not understand; for instance, if the system would analyze and display what (mis)conceptions a student is likely to have, based on their answers. More advanced dashboards cannot only help teachers to monitor students' performance but also with making instructional decisions: They can *alert* teachers to events that need attention (e.g., a poor performing student) and could potentially also *advise* teachers on the type of help a student needs (cf. Holstein et al., 2019, who investigated the different roles the dashboards could potentially fulfill in helping teachers to support student learning).

Finally, the online learning systems could potentially help teachers to identify the students who are most in need of support with developing their monitoring and regulation skills, which teachers seem to find difficult (cf. Chapter 6). For instance, when learning systems ask students to self-score, and decide on a next activity, teacher dashboards could collect and share the information on students' monitoring and regulation (accuracy) in the teacher dashboards.



## 7.5. Conclusion

The research presented in this dissertation provided new insights into primary school students' and teachers' monitoring and regulation accuracy of students' mathematics problem-solving performance, and on how their accuracy can be improved. The studies presented in this dissertation support the idea that having students self-score their answers provides them with more diagnostic performance cues, which improves their monitoring accuracy and (to some extent) regulation accuracy, as well as their awareness of their regulation (in)accuracy, which is important for improving the effectiveness of students' self-regulated learning. Similarly, providing teachers with access to (easy-to-infer) performance cues, improves their monitoring of their students' performance, as well as their accuracy awareness, which is important for their ability to provide adaptive instruction. Yet, when it comes to developing students' regulation skills, both students and teachers seem to need further support: Even after self-scoring, many students still need help from their teachers to make accurate regulation decisions, while teachers have difficulty identifying which students need their help in doing so. These findings set the stage for further (practice-oriented) research on how cue use can influence students' and teachers' monitoring (awareness) and in turn their regulation accuracy (awareness), and for developing and testing interventions to improve teachers' differentiation practices and students' self-regulated learning skills.



## Nederlandstalige Samenvatting

In het basisonderwijs zijn leerkrachten en leerlingen samen verantwoordelijk voor het leerproces van de leerling. Aan de ene kant wordt van leerkrachten verwacht dat ze gedifferentieerd lesgeven, wat betekent dat ze hun onderwijs aanpassen aan de verschillende behoeften van individuele leerlingen (Parsons et al., 2018; Tomlinson et al., 2003; Van de Pol et al., 2010). Aan de andere kant wordt van de leerlingen in toenemende mate verwacht dat ze zelf hun leerproces (leren te) sturen (dit wordt ook wel 'leren leren' genoemd), omdat dit de basis vormt voor een leven lang leren (Bjork et al., 2013; OECD, 2022) en leerlingen die goed zijn in het zelfgestuurd leren ook betere leerresultaten laten zien (Dent & Koenka, 2016). Daarom wordt van leerkrachten ook steeds meer verwacht dat ze hun leerlingen helpen bij het ontwikkelen van de vaardigheden die nodig zijn om zelf hun leerproces goed te kunnen sturen (Dignath & Büttner, 2018).

Twee centrale processen in zowel zelfgestuurd leren (De Bruin & Van Gog, 2012; Griffin et al., 2013) als gedifferentieerd lesgeven (Shavelson & Stern, 1981; Thiede et al., 2019; Van de Pol et al., 2014) zijn monitoring en regulatie. Monitoring is het evalueren van het huidige prestatieniveau van individuele leerlingen—dat wil zeggen, inschattingen maken van wat leerlingen wel en niet beheersen met betrekking tot een bepaalde taak. Regulatie betreft het maken van beslissingen over welke vervolgvactiteiten (bijv. iets nog eens bestuderen, extra oefenen, extra instructie) leerlingen het beste kunnen uitvoeren om hun prestatieniveau te verbeteren. Accurate monitoring is een essentiële (maar niet de enige) voorwaarde voor accurate regulatie – zonder een goede inschatting van de huidige prestatie is het immers niet mogelijk een vervolgvactiteit te kiezen die past bij de eigenlijke behoefte van de leerling (Dunlosky & Rawson, 2012; Metcalfe & Finn, 2008; Pintrich, 2000; Winne & Hadwin 1998; Zimmerman, 2000).

Helaas heeft eerder onderzoek laten zien dat basisschoolleerlingen vaak inaccurate monitoring-inschattingen (en regulatie-beslissingen) maken, bijvoorbeeld tijdens het uit het hoofd leren van informatie (e.g., Roebers et al., 2014), begrijpend lezen (e.g., De Bruin et al., 2011) en rekenen/wiskunde (e.g., Baars et al., 2014a; Boekaerts & Rozendaal, 2010; García et al., 2016). Ook leerkrachten maken vaak inaccurate monitoring-inschattingen van leerlingprestaties, bijvoorbeeld op het gebied van woordenschat, begrijpend lezen, en rekenen/wiskunde (voor reviewstudies, zie Kaufman, 2020; Südkamp et al., 2012; Urhahne & Wijnia, 2021). Dat is een probleem, aangezien accurate monitoring essentieel is voor accurate regulatie (voor leerkrachten: Van de Pol et al., 2011; voor leerlingen: Dunlosky & Rawson, 2012), dus daardoor zullen de beslissingen van leerkrachten en leerlingen t.a.v. vervolgvactiteiten niet altijd optimaal aansluiten bij de behoeften van de leerlingen. Dat leidt er vervolgens toe dat leerlingen onnodig tijd besteden aan iets wat ze al kunnen, of juist met een veel te moeilijke taak geconfronteerd worden; beide

gevallen hebben niet alleen negatieve gevolgen voor leerresultaten, maar ook voor de motivatie (Seegers & Boekaerts, 1993).

Een verklaring voor het feit dat monitoring-inschattingen vaak inaccuraat zijn, is dat zowel leerlingen (Koriat, 1997) als leerkrachten (Byers & Evans, 1980; Snow, 1968) allerlei verschillende soorten informatie of *cues* gebruiken als ze het huidige prestatieniveau van zichzelf of hun leerlingen inschatten. Des te voorspellender of *diagnostischer* de gebruikte cues zijn voor de eigenlijke prestatie van de leerling (d.w.z. des te sterker de correlatie tussen de cue en de prestatie), des te accurater de inschattingen van de leerling of leerkracht zullen zijn (Koriat, 1997; Thiede et al., 2010, 2015; Van Loon et al., 2014). Bijvoorbeeld: als een leerkracht of leerling een inschatting maakt van hoe een leerling ervoor staat op een bepaalde taak, is de prestatie op een eerder gemaakte soortgelijke taak waarschijnlijk een meer diagnostische cue dan de interesse van de leerling in de taak of de tijd die een leerling besteed heeft aan de taak. Inaccurate monitoring komt tot stand doordat leerlingen en leerkrachten niet altijd cues met een hoge diagnosticiteit gebruiken. Dus, om zelfgestuurd leren en gedifferentieerd lesgeven te kunnen verbeteren, is het van belang om te onderzoeken of interventies die leerkrachten en leerlingen toegang geven tot meer diagnostische cues leiden tot verbeteringen in de accuratesse van hun monitoring-inschattingen en de daaropvolgende regulatie-beslissingen.

Het hoofddoel van dit proefschrift was om meer inzicht te krijgen in hoe de accuratesse van de monitoring-inschattingen en regulatie-beslissingen die leerlingen en leerkrachten in het basisonderwijs maken van de rekenprestaties van de leerling, wordt beïnvloed door interventies die ervoor zorgen dat ze op meer diagnostische cues focussen. Om precies te zijn is onderzocht of, en zo ja hoe, bij *leerlingen* het zelf nakijken van hun antwoorden de monitoring- en regulatie-accuratesse, en hun bewustzijn van de (in)accuratesse, zou verbeteren; of, en zo ja hoe, bij *leerkrachten* de accuratesse van de monitoring-inschattingen van de rekenprestaties van hun leerlingen, en hun bewustzijn van die (in)accuratesse, zou verbeteren wanneer ze worden voorzien van informatie over eerdere prestaties van de leerlingen; en hoe goed leerkrachten de monitoring- en regulatie-accuratesse van hun leerlingen in zouden kunnen schatten.

In de hoofdstukken 2 t/m 6 van dit proefschrift staan vijf empirische studies beschreven, uitgevoerd in groep 6 (de leerlingen zijn dan 9-10 jaar oud). De data voor de studies die staan beschreven in de hoofdstukken 2, 3, 5 en 6 zijn verzameld tijdens één grote dataverzameling; hoofdstuk 4 betreft een aparte dataverzameling.

## Onderzoeksresultaten

De studies in **hoofdstuk 2 en 3** hadden betrekking op het verbeteren van de monitoring- en regulatie-accuratesse (en hun bewustzijn daarvan) bij *leerlingen*. De studie in **hoofdstuk 2** onderzocht of zelf nakijken zou leiden tot verbeterde monitoring- en regulatie-accuratesse van leerlingen tijdens het maken van procedurele rekentaken (vermenigvuldigen en delen). Met zelf nakijken wordt bedoeld dat leerlingen hun zelf gegeven antwoorden vergelijken met de juiste antwoorden, wat leerlingen toegang geeft tot *prestatieclues* met een hoge diagnosticiteit (Rawson & Dunlosky, 2007). Bovendien werd onderzocht of de effecten van zelf nakijken op (potentiële) verbeteringen in de monitoring- en regulatie-accuratesse zouden verschillen tussen laag en hoog presterende leerlingen, aangezien dit zou vragen om verschillende benaderingen in interventies. De absolute monitoring-accuratesse van de leerlingen werd gemeten door het absolute verschil (d.w.z. ongeacht of het positief of negatief was) te berekenen tussen de inschattingen van leerlingen van hoeveel opgaven zij correct hadden beantwoord en het aantal opgaven dat zij daadwerkelijk correct hadden beantwoord. De absolute regulatie-accuratesse werd bepaald door het absolute verschil te berekenen tussen de behoefte aan extra oefening of extra instructie zoals ingeschat door de leerling en zoals deze daadwerkelijk was (zoals bepaald door experts) met betrekking tot de vermenigvuldigings- of deeltaak (d.w.z. de taak m.b.t. de vaardigheid delen).

De resultaten lieten zien dat het zelf nakijken van antwoorden op procedurele rekenproblemen een effectieve manier is om de monitoring-accuratesse te verbeteren (de meeste leerlingen maakten perfect accurate monitoring-inschattingen na het nakijken). Dit vertaalde zich ook grotendeels in accuratere regulatie-beslissingen (d.w.z. in de hele groep van deelnemende leerlingen op beide taken en in de groepen van de laag en hoog presterende leerlingen op de vermenigvuldigingstaak). Met betrekking tot monitoring maakten de hoog presterende leerlingen vóór het nakijken accuratere inschattingen dan laag presterende leerlingen. Echter, na het nakijken verdween dit verschil op de deeltaak en werd het verschil heel klein (maar het bleef significant) op de vermenigvuldigingstaak. Op het gebied van regulatie maakten hoog presterende leerlingen substantieel accuratere beslissingen dan laag presterende leerlingen, voor en na het nakijken, op de vermenigvuldigingstaak (niet op de deeltaak). Kortom, zelf nakijken is een effectieve manier om de accuratesse van het zelf monitoren en reguleren tijdens het maken van procedurele rekentaken te verbeteren. Er is echter een grote groep leerlingen (vooral de laag presterende leerlingen) die ook na zelf te hebben nagekeken nog steeds hulp van hun leerkracht nodig heeft bij het maken van accurate regulatie-beslissingen.

Echter, alleen de monitoring- en regulatie-accuratesse verhogen is mogelijk niet voldoende om het zelfgestuurd leren daadwerkelijk te verbeteren; het is vermoedelijk ook van belang dat leerlingen zich bewust zijn van de accuratesse van hun inschattingen. Dit

omdat het *accuratesse-bewustzijn* vermoedelijk bepaalt of (en hoe) leerlingen handelen op basis van hun inschattingen (zoals ook gesuggereerd door Gabriele et al., 2016; Händel & Fritzsche, 2016; Patterson et al., 2001). In de studie in **hoofdstuk 3** werd daarom geëxploreerd of leerlingen zich bewust waren van hun monitoring- en regulatie-accuratesse en of, en zo ja hoe, dit werd beïnvloed door het zelf nakijken van hun antwoorden.

De leerlingen werd gevraagd om, na het maken van de monitoring-inschatting of regulatie-beslissing, aan te geven hoeveel vertrouwen ze erin hadden dat hun monitoring-inschatting (voor het nakijken) en regulatie-beslissing (voor en na het nakijken) accuraat zouden zijn. Tijdens een pilotstudie werd de leerlingen ook gevraagd naar hun vertrouwen in hun monitoring-accuratesse ná het nakijken, maar omdat alle leerlingen daar maximaal vertrouwen in hadden (en de vraag "raar" vonden) is deze vraag weer verwijderd.

Leerlingen tonen bewustzijn van hun (in)accuratesse wanneer zich zekerder voelen over accuratere inschattingen dan over minder accurate inschattingen en vice versa. De resultaten wezen uit dat de leerlingen zich gemiddeld genomen zeer beperkt bewust waren van de (in)accuratesse van hun monitoring-inschattingen en regulatie-beslissingen die ze maakten voordat ze hun antwoorden hadden nagekeken. Over het algemeen (d.w.z. in de gehele groep van deelnemende leerlingen), leek zelf nakijken een positief effect te hebben op het bewustzijn van hun regulatie-accuratesse bij de leerlingen. Dit effect was echter minder duidelijk aanwezig onder de laag presterende leerlingen en leerlingen van wie de regulatie-beslissingen minder accuraat werden of inaccuraat bleven nadat ze hun antwoorden hadden nagekeken. Toekomstig onderzoek zal moeten uitwijzen of er aanvullende of andere manieren (zoals feedback of training) mogelijk zijn om het bewustzijn van de regulatie-accuratesse te verbeteren.

De studies in **hoofdstuk 4, 5 en 6** hadden betrekking op leerkrachten en onderzochten het verbeteren van (hun bewustzijn van) hun inschattingen van de leerlingprestaties, en hun inschattingen van de monitoring- en regulatie-accuratesse van de leerlingen. In de studies in hoofdstuk 4 en 5 werd gemanipuleerd of prestatie cues (d.w.z. de prestaties van leerlingen op eerder gemaakte rekentaken) en/of leerling cues (d.w.z. algemene leerlingkenmerken zoals hun interesse in rekenen of hun nationaliteit) al dan niet beschikbaar waren. Naar verwachting zou het beschikbaar stellen van informatie over de prestaties van leerlingen (antwoorden of scores op eerdere, vergelijkbare oefentaken) waaruit meer diagnostische prestatie cues afgeleid zouden kunnen worden, het cue-gebruik van leerkrachten beïnvloeden en hun monitoring-inschattingen van de leerlingprestaties verbeteren, in vergelijking met een situatie waarin ze alleen over leerling cues beschikken (o.b.v. hun kennis over de leerling).

In de studie in **hoofdstuk 4** maakten de leerkrachten inschattingen over het conceptuele begrip van de grootte van kommagetallen bij hun leerlingen. De leerlingen maakten een oefentaak en een vervolgtak, terwijl leerkrachten alleen inschattingen maakten over de leerlingprestaties op de vervolgtak. Iedere leerkracht maakte deze

inschattingen in drie condities, waarbij ze werden voorzien van: (1) alleen de namen van de leerlingen over wie ze inschattingen maakten (d.w.z. alleen leerlingcues beschikbaar), (2) alleen geanonimiseerde antwoorden van de leerlingen op de oefenopgaven over kommagetallen (d.w.z. alleen prestatiecues beschikbaar), (3) zowel de namen van de leerlingen als hun antwoorden (d.w.z. zowel leerling- als prestatiecues beschikbaar). Via de namen van de leerlingen hadden de leerkrachten toegang tot leerlingcues, gebaseerd op wat ze over de leerlingen wisten. Naar verwachting waren de leerlingcues niet of beperkt diagnostisch voor hoe leerlingen presteerden op opgaven over de grootte van kommagetallen. Uit (foutpatronen in) de antwoorden van de leerlingen op de oefenopgaven konden leerkrachten de (mis)concepties afleiden die de leerlingen met betrekking tot kommagetallen hadden, waarmee ze toegang hadden tot prestatiescues met een hoge diagnosticiteit.

Leerkrachten maakten opgave-specifieke inschattingen over het conceptuele begrip van kommagetallen van hun leerlingen, door in te schatten hoe leerlingen zouden presteren per opgave van de vervolftaak. De accuratesse van deze leerkrachtinschattingen werd gemeten in termen van sensitiviteit (accuraat inschatten wat leerlingen wel begrijpen) en specificiteit (accuraat inschatten wat leerlingen niet begrijpen). De leerkrachten werd gevraagd om hardop na te denken terwijl ze de inschattingen maakten, om hun cue-gebruik te meten.

Uit de resultaten bleek dat wanneer leerkrachten werden voorzien van de antwoorden van de leerlingen op de oefentaak, naast de namen van de leerlingen, dit niet leidde tot significante verbeteringen in de sensitiviteit of specificiteit van de leerkrachtinschattingen van het begrip van de grootte van kommagetallen bij hun leerlingen. Een mogelijke verklaring hiervoor is dat de leerkrachten – ondanks de aanwezigheid van prestatiecues – *niet minder* gebruik gingen maken van (de weinig diagnostische) leerlingcues; ze gebruikten deze nog evenveel als wanneer ze alleen werden voorzien van de leerlingnamen, zoals bleek uit de analyses van de hardop-denken-protocollen. Wanneer de leerkrachten alleen toegang hadden tot de antwoorden van de leerlingen op oefentaken (d.w.z. zonder de leerlingnamen, waardoor ze geen toegang hadden tot leerlingcues), verbeterde wel de specificiteit, maar niet de sensitiviteit van de leerkrachtinschattingen.

Het was voor de leerkrachten in de studie in hoofdstuk 4 echter ook niet eenvoudig om de (mis)concepties uit de antwoorden van de leerlingen af te leiden. Wanneer er informatie beschikbaar is die niet zo moeilijk te interpreteren is, zou het voor leerkrachten wellicht gemakkelijker kunnen zijn om hun inschattingen op de meer diagnostische prestatiecues te baseren en tegelijkertijd de minder diagnostische leerlingcues te negeren. Dit zou kunnen leiden tot accuratere monitoring-inschattingen.

De studie in **hoofdstuk 5** testte deze hypothese, door leerkrachten te voorzien van prestatiecues die niet om interpretatie vragen, namelijk de scores van leerlingen op eerder gemaakte taken. Leerkrachten maakten inschattingen van hoe hun leerlingen

zouden presteren op een vermenigvuldigings- en deeltaak. Iedere leerkracht deed dit onder twee condities: terwijl ze waren voorzien van (1) alleen de namen van de leerlingen (d.w.z. alleen leerlingcues beschikbaar) en (2) zowel de namen van de leerlingen als hun scores op vergelijkbare vermenigvuldigings- en deeltaken, die ze een week eerder hadden gemaakt (d.w.z. leerling- en prestatiecues beschikbaar). De absolute monitoring-accuratesse van de leerkrachten werd gemeten door het absolute verschil te berekenen tussen de inschatting van hoeveel opgaven een leerling correct had beantwoord op de vermenigvuldigings- of deeltaak en de feitelijke prestatie van de leerling op die taak. Een indicatie van het cue-gebruik van de leerkracht werd gemeten door het berekenen van correlaties tussen de monitoring-inschattingen van de leerkrachten en de leerkrachtpercepties van algemene leerlingkenmerken (gemeten d.m.v. een vragenlijst met items over bijv. interesse in rekenen en nationaliteit).

Net als voor leerlingen is het voor leerkrachten belangrijk dat ze, naast dat ze accurate monitoring-inschattingen maken, zich bewust zijn van de (in)accuratesse van hun inschattingen. Dit omdat leerkrachten, wanneer ze zich bewust zijn van accurate inschattingen (d.w.z. vertrouwen hebben in die inschattingen), ze vermoedelijk passende instructionele beslissingen nemen gebaseerd op die accurate inschattingen; Wanneer ze zich bewust zijn van inaccurate inschattingen (d.w.z. geen vertrouwen in die inschattingen hebben), kunnen ze op zoek gaan naar meer informatie over het prestatieniveau van de leerling, voordat ze een instructionele beslissingen nemen (Gabriele et al., 2016).

Daarom werd ook geëxploreerd in hoeverre leerkrachten zich bewust zouden zijn van de (in)accuratesse van hun monitoring-inschattingen en of, en zo ja hoe, dit werd beïnvloed door de beschikbaarheid van prestatiecues. Om dit type bewustzijn te meten werd de leerkrachten gevraagd om aan te geven hoeveel vertrouwen ze erin hadden dat hun monitoring-inschattingen accuraat zouden zijn.

De bevindingen in hoofdstuk 5 lieten zien dat wanneer leerkrachten werden voorzien van de scores van leerlingen op eerder gemaakte taken, naast de namen van de leerlingen, hun absolute monitoring-accuratesse van de leerlingprestaties op de keer- en deeltaak verbeterde. De (correlationele) analyses van het cue-gebruik van de leerkrachten suggereren dat dit effect tot stand kwam doordat leerkrachten niet alleen de prestatiecues gebruikten, maar ook de minder diagnostische leerlingcues negeerden, wanneer de scores op eerdere gemaakte taken beschikbaar waren. Verder bleek dat wanneer leerkrachten alleen toegang hadden tot leerlingcues, ze zich al enigszins bewust waren van hun (in)accuratesse, in die zin dat ze meer vertrouwen hadden in accuratere inschattingen en minder vertrouwen in minder accurate inschattingen. Het vertrouwen in, en het bewustzijn van, hun (in)accuratesse namen verder toe wanneer de leerkrachten werden voorzien van de scores op de eerder gemaakte taken.

Kortom, de bevindingen uit hoofdstuk 4 en 5 suggereren dat de monitoring-inschattingen van de leerkrachten verbeteren als ze worden voorzien van prestatiecues,



maar alleen als deze cues makkelijk af te leiden zijn uit leerlingwerk en/of het ze lukt om leerlingcues te negeren (leerlingcues zijn in de onderwijspraktijk altijd aanwezig; leerkrachten kunnen hun kennis over hun leerlingen niet even uitschakelen). Vervolgonderzoek zou deze conclusie kunnen verifiëren, bijvoorbeeld door onderzoek te doen naar het effect van moeilijk versus gemakkelijk af te leiden prestatie cues, binnen één en dezelfde studie.

In de studie in **hoofdstuk 6** werd geëxploreerd hoe goed leerkrachten de monitoring- en regulatie-vaardigheden van hun leerlingen in zouden kunnen schatten. Het is belangrijk dat deze inschattingen accuraat zijn, zodat leerkrachten de leerlingen kunnen identificeren die het meest inaccuraat zijn en dus de meeste hulp nodig hebben bij het ontwikkelen van hun monitoring- en regulatie-vaardigheden. Om dit te kunnen onderzoeken maakten leerkrachten niet alleen inschattingen van de prestaties en behoefte aan extra oefening of instructie van hun leerlingen, maar gaven ze ook aan welke monitoring-inschattingen en regulatie-beslissingen ze dachten dat hun leerlingen hadden gemaakt. Op basis hiervan werd de accuratesse van de leerkrachtinschattingen van de monitoring- en regulatie-accuratesse van hun leerlingen berekend. Bovendien werd geëxploreerd of, en zo ja hoe, de accuratesse van de leerkrachtinschattingen van de monitoring- en regulatie-accuratesse van hun leerlingen beïnvloed zouden worden door de leerkrachtpercepties van algemene leerlingkenmerken. Op deze manier kon meer inzicht worden verkregen in hoe (in)accurate leerkrachtinschattingen van de monitoring- en regulatie-vaardigheden van hun leerlingen tot stand komen.

Uit de resultaten bleek dat de leerkrachten vrij goed wisten dat hun leerlingen over het algemeen redelijk accurate monitoring-inschattingen en regulatie-beslissingen hadden gemaakt. De leerkrachten hadden echter moeite met het identificeren van leerlingen die behoorlijk inaccuraat waren (voor wie het extra belangrijk is dat ze hulp krijgen bij het monitoren en reguleren). Samengenomen verklaarden de leerkrachtpercepties van de verschillende leerlingkenmerken een aanzienlijk deel van de variantie in de accuratesse van de leerkrachtinschattingen van de monitoring- en regulatie-accuratesse van hun leerlingen. Dit suggereert dat het algemene beeld dat leerkrachten van hun leerlingen hebben mogelijk de (accuratesse van) deze leerkrachtinschattingen beïnvloedt. Kijkend naar de invloed van de specifieke leerlingkenmerken, bleek dat de leerkrachtinschattingen van de regulatie-accuratesse van hun leerlingen voordat ze hun antwoorden hadden nagekeken onterecht werden beïnvloed door de nationaliteit van de leerlingen: de inschattingen van de leerkracht waren gemiddeld genomen accurater als de leerling een meer westerse achtergrond had. Dus, deze bevindingen laten zien dat leerkrachten waarschijnlijk hulp kunnen gebruiken bij het in beeld brengen van welke leerlingen substantieel inaccurate monitoring-inschattingen en regulatie-beslissingen maken, alsook bij het tegengaan van potentiële vooroordelen bij het maken van die inschattingen, om hun leerlingen zo optimaal te kunnen helpen bij het ontwikkelen van monitoring- en regulatie-vaardigheden.

## Conclusies en Aanbevelingen voor de Onderwijspraktijk

De studies uit dit proefschrift geven inzicht in hoe de accuratesse van de monitoring-inschattingen en regulatie-beslissingen die leerlingen en leerkrachten in het basisonderwijs maken van de rekenprestaties van de leerling verbeterd zouden kunnen worden. Ten eerste ondersteunt het onderzoek in dit proefschrift het idee dat zelf nakijken, zoals dat in de bovenbouw van de meeste basisscholen veel gebeurt, kan bijdragen aan effectiever zelfgestuurd leren: zelf nakijken leidde namelijk tot verbeteringen in de monitoring-accuratesse en – in zekere mate – in de regulatie-accuratesse van de leerlingen, alsmede in hun bewustzijn van de regulatie-accuratesse. Echter, er is een grote groep leerlingen (vooral de laag presterende leerlingen) die ook na zelf te hebben nagekeken nog steeds hulp nodig heeft bij het maken van accurate regulatie-beslissingen. Daarom is het van belang dat leerkrachten er niet vanuit gaan dat leerlingen hun vervolgbehoeften (bijv. extra instructie) goed kunnen inschatten nadat ze hun werk zelf hebben nagekeken en daarnaast, dat ze hun leerlingen expliciete instructie geven over hoe ze deze beslissingen moeten maken (iets wat leerkrachten maar zelden blijken te doen; Dignath & Veenman, 2021). Toekomstig praktijkgericht onderzoek zal antwoord moeten geven op de vraag hoe leerkrachten ondersteund kunnen worden bij het expliciet onderwijzen van regulatie-vaardigheden en bij het identificeren van welke leerlingen hier het meest behoefte aan hebben (aangezien uit de bevindingen in hoofdstuk 6 bleek dat dit lastig was voor leerkrachten). Deze kennis kan vervolgens geïmplementeerd worden in nascholingscursussen en lerarenopleidingen.

Daarnaast ondersteunt het onderzoek in dit proefschrift het idee dat leerkrachtinschattingen van de rekenprestaties van hun leerlingen, en hun accuratessebewustzijn, verbeteren wanneer ze worden voorzien van informatie waaruit ze gemakkelijk prestatieclues kunnen afleiden. Dit is van belang om effectiever te kunnen differentiëren. Verder onderzoek zal moeten uitwijzen wat voor leerlingwerk of interventies het meest geschikt zijn om leerkrachten te voorzien van gemakkelijk af te leiden prestatieclues; hoe leerkrachten geholpen kunnen worden om instructionele beslissingen op deze cues te baseren zonder dat het extra tijd kost; en hoe leerkrachten geholpen kunnen worden bij het negeren van minder diagnostische leerlingclues.

Mogelijk kan het gebruik van online oefensystemen dit alles (in de toekomst) vergemakkelijken. Veel online oefensystemen die momenteel in scholen gebruikt worden, kunnen leerkrachten al ondersteunen bij het nemen van instructionele beslissingen die optimaal aansluiten bij de behoefte van hun leerlingen, door overzichten te geven van prestatieclues per leerling en hierbij uit te lichten welke leerlingen ondermaats presteren. Potentieel kunnen deze systemen ook een rol spelen in het bevorderen van zelf-monitoring en zelf-regulatie van leerlingen (en hun bewustzijn hiervan), bijvoorbeeld door

het geven van feedback op de accuratesse (cf. Roll et al., 2011; Molenaar et al., 2020). Dat zou een uitkomst zijn, want diagnostische prestatie cues verzamelen om optimaal te kunnen differentiëren én leerlingen helpen bij het ontwikkelen van de vaardigheden om hun leerproces zelf te sturen vraagt veel van leerkrachten, die toch al kampen met een hoge werkdruk (Gemmink et al., 2020).

Kortom, deze bevindingen vormen een basis voor verder (praktijkgericht) onderzoek naar hoe cue-gebruik de monitoring-accuratesse en regulatie-accuratesse van leerlingen en leerkrachten (evenals hun bewustzijn hiervan) kan beïnvloeden, alsmede voor het ontwikkelen van interventies ter verbetering van gedifferentieerd lesgeven en zelfgestuurd leren.



# Supplementary Materials - Chapter 2

## 2.1. Coding Students' Need for Intervention

**Table S2.1**

*Examples of Procedural and Computational Errors*

Type of Error	Example when problem is 6 x 472	Example when problem is 228 : 3
Use of the wrong strategy or lack of use of a specific strategy (procedural error).	$\begin{array}{r} 2400 \\ 420 \\ \underline{12} + \\ 7800 \end{array}$	$\begin{array}{r} 228 : 3 = \\ 200 \quad 28 \end{array}$
Wrong use of a correct strategy (procedural error).	$\begin{array}{r} \textcircled{4} \textcircled{1} \\ 472 \\ \underline{\quad 6} \times \\ 2422 \end{array}$	$\begin{array}{r} 3/228 \setminus 742 \\ 21 - \\ \underline{12} \\ 08 \\ \underline{\quad 6} - \\ 2 \end{array}$
Computational error.	$\begin{array}{l} 6 \times 2 = 10 \\ 6 \times 70 = 480 \end{array}$	$\begin{array}{l} 210 : 3 = 80 \\ 18 : 3 = 7 \end{array}$

## 2.2. Scoring Students' Regulation Accuracy

**Table S2.2**

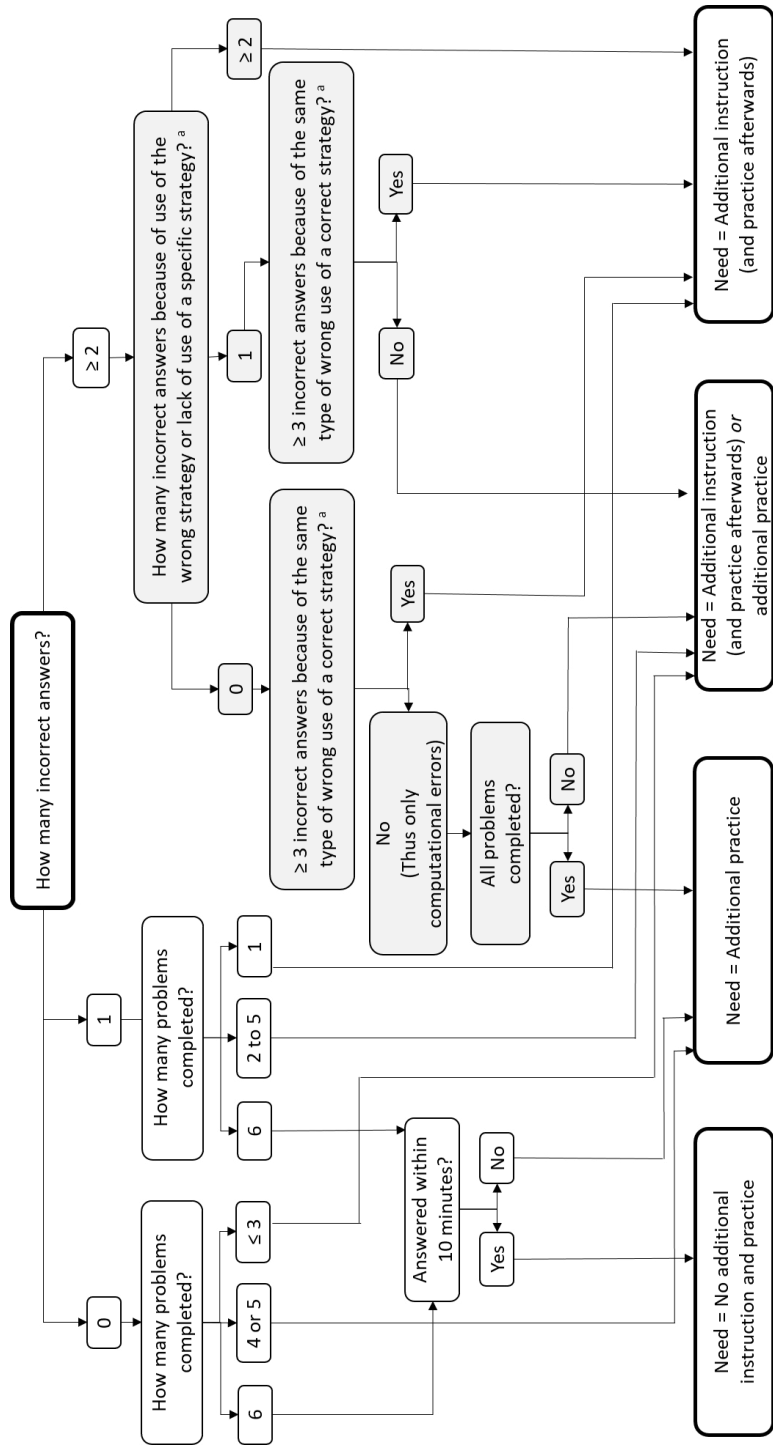
*Cross Tabulation of Scoring Students' Regulation Accuracy*

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

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

**Figure S2.1**

*Coding Scheme for Actual Needs of Intervention*



Note. Unshaded parts were code automatically, shaded parts were coded manually, <sup>a</sup>See Table S2.1 for examples. A question mark/cross/line was coded as "lack of strategy".

## 2.3. Supplementary Results

**Table S2.3**

*The Comparison of Low- and High-Performing Students' Absolute Monitoring/Regulation Accuracy Before Self-Scoring (RQ 1A and 1B)*

	Absolute monitoring accuracy before self-scoring			Absolute regulation accuracy before self-scoring		
	Multiplication, <i>n</i> = 597			Division, <i>n</i> = 374		
	M0: intercept only	M1: effect of performance	M0: intercept only	M1: effect of performance	M0: intercept only	M1: effect of performance
	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )
Fixed effects						
Intercept	1.21 (0.05)***	1.67 (0.10)***	0.89 (0.06)***	1.04 (0.10)***	0.50 (0.03)***	0.63 (0.04)***
Low/high <sup>a</sup>		-0.84 (0.12)***		-0.30 (0.11)**		-0.26 (0.05)***
Random effects						
$\sigma^2_e$ (day)	1.29 (0.19)***	1.24 (0.18)***	0.59 (0.06)***	0.59 (0.06)***	0.34 (0.03)***	0.34 (0.03)***
$\sigma^2_{u0}$ (student)	0.16 (0.09)	0.04 (0.09)	0.16 (0.08)*	0.14 (0.07)*	0.09 (0.02)***	0.07 (0.02)**

Note. <sup>a</sup>0 = low-performing students; 1 = high-performing students

\*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

**Table S2.4**

*The Effect of Self-Scoring on Absolute Monitoring/Regulation Accuracy, Within the Whole Sample (RQ 2A and 2B)*

	Absolute monitoring Accuracy			Absolute regulation Accuracy		
	Division, N = 1958			Division, N = 1362		
	Multiplication, N = 1389			Multiplication, N = 1927		
	M0: intercept only	M1: effect of self-scoring	M0: intercept only	M1: effect of self-scoring	M0: intercept only	M1: effect of self-scoring
	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)
Fixed effects						
Intercept	0.70 (0.02)***	1.23 (0.04)***	0.54 (0.03)***	0.95 (0.05)***	0.53 (0.02)***	0.61 (0.02)***
Self-Scoring <sup>a</sup>		-1.08 (0.04)***		-0.84 (0.05)***		-0.17 (0.02)***
Random effects						
$\sigma^2_e$ (self-scoring)	1.10 (0.06)***	0.73 (0.06)***	0.68 (0.05)***	0.46 (0.03)***	0.21 (0.02)***	0.20 (0.02)***
$\sigma^2_{u0}$ (day)	0.00 (0.09) <sup>b</sup>	0.01 (0.03)	0.00 (0.10) <sup>c</sup>	0.01 (0.02)	0.15 (0.02)***	0.16 (0.02)***
$\sigma^2_{v0}$ (student)	0.00 (0.05) <sup>b</sup>	0.07 (0.03) <sup>*</sup>	0.00 (0.05) <sup>c</sup>	0.04 (0.02) <sup>*</sup>	0.08 (0.02)***	0.08 (0.02)***
					0.05 (0.02) <sup>*</sup>	0.05 (0.02) <sup>*</sup>

Note. <sup>a</sup>0 = before self-scoring, 1 = after self-scoring; Unrounded values: <sup>b</sup> $\sigma^2_{u0} = 0.004$ , <sup>c</sup> $\sigma^2_{v0} = 0.004$ ,  $\sigma^2_e = 0.002$ ,  $\sigma^2_{v0} = 0.003$ .  
 \*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$



**Table S2.5**

*The Effect of Self-Scoring on Absolute Monitoring/Regulation Accuracy, Within Low-Performing Students (RQ 2A and 2B)*

	Absolute monitoring Accuracy				Absolute regulation Accuracy			
	Multiplication, n = 535		Division, n = 386		Multiplication, n = 549		Division, n = 378	
	M0: intercept only	M1: effect of self-scoring	M0: intercept only	M1: effect of self-scoring	M0: intercept only	M1: effect of self-scoring	M0: intercept only	M1: effect of self-scoring
	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)
Fixed effects								
Intercept	0.90 (0.05)***	1.58 (0.10) ***	0.59 (0.05)***	1.04 (0.10)***	0.56 (0.03)***	0.63 (0.04)***	0.28 (0.05)***	0.31 (0.05)***
Self-Scoring <sup>a</sup>		-1.37 (0.08)***		-0.93 (0.12)***		-0.16 (0.04)***		-0.05 (0.03)
Random effects								
$\sigma^2_e$ (self-scoring)	1.45 (0.11)***	0.86 (0.09)***	0.77 (0.10)***	0.50 (0.06)***	0.25 (0.03)***	0.24 (0.03)***	0.06 (0.01)*** <sup>d</sup>	0.06 (0.01)*** <sup>e</sup>
$\sigma^2_{u0}$ (day)	0.01 (0.24)	0.08 (0.08)	0.00 (0.02) <sup>b</sup>	0.00 (0.04) <sup>c</sup>	0.15 (0.04)***	0.16 (0.04)***	0.11 (0.04)*** <sup>d</sup>	0.11 (0.04)*** <sup>e</sup>
$\sigma^2_{v0}$ (student)	0.01 (0.14)	0.06 (0.06)	0.01 (0.04)	0.07 (0.05)	0.05 (0.03)	0.05 (0.03)	0.10 (0.06) <sup>d</sup>	0.10 (0.06) <sup>e</sup>

Note. <sup>a</sup> 0 = before self-scoring, 1 = after self-scoring; Unrounded values; <sup>b</sup>  $\sigma^2_{u0} = 0.003$ ; <sup>c</sup>  $\sigma^2_{u0} = 0.004$ ; Effects with outliers: <sup>d</sup>  $\sigma^2_e = 0.13$  (0.03)\*\*\*,  $\sigma^2_{u0} = 0.08$  (0.04)\*,  $\sigma^2_{v0} = 0.10$  (0.05)\*; <sup>e</sup>  $\sigma^2_e = 0.13$  (0.03)\*\*\*,  $\sigma^2_{u0} = 0.10$  (0.05)\*

\*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

**Table S2.6**

*The Effect of Self-Scoring on Absolute Monitoring/Regulation Accuracy, Within High-Performing Students (RQ 2A and 2B)*

	Absolute monitoring Accuracy			Absolute regulation Accuracy		
	Multiplication, <i>n</i> = 634	Division, <i>n</i> = 403	Multiplication, <i>n</i> = 604	Division, <i>n</i> = 370		
	M0: intercept only	M0: intercept only	M0: intercept only	M0: intercept only	M1: effect of self-scoring	M1: effect of self-scoring
	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )
Fixed effects						
Intercept	0.43 (0.02)***	0.39 (0.03)***	0.29 (0.02)***	0.18 (0.02)***	0.33 (0.03)***	0.19 (0.03)***
Self-Scoring <sup>a</sup>	-0.71 (0.04)***	-0.59 (0.05)***	-0.09 (0.02)***			-0.02 (0.03)
Random effects						
$\sigma^2_e$ (self-scoring)	0.41 (0.03)***	0.27 (0.02)***	0.07 (0.01)***	0.07 (0.01)***	0.07 (0.01)***	0.07 (0.01)***
$\sigma^2_{u0}$ (day)	0.00 (0.04) <sup>b</sup>	0.00 (0.03) <sup>c</sup>	0.16 (0.02)***	0.07 (0.02)**	0.16 (0.02)***	0.07 (0.02)**
$\sigma^2_{v0}$ (student)	0.00 (0.04) <sup>b</sup>	0.00 (0.03) <sup>c</sup>	0.02 (0.02) <sup>d</sup>	0.01 (0.01)	0.01 (0.02)	0.01 (0.01)

Note. <sup>a</sup> 0 = before self-scoring, 1 = after self-scoring; Unrounded values: <sup>b</sup>  $\sigma^2_{u0} = 0.001$ ,  $\sigma^2_{v0} = 0.001$ ; <sup>c</sup>  $\sigma^2_{u0} = 0.002$ ,  $\sigma^2_{v0} = 0.001$ ; <sup>d</sup>  $\sigma^2_{v0} = 0.001$   
 \*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

Table S2.7

The Comparison of Low- and High-Performing Students' Absolute Monitoring/Regulation Accuracy After Self-Scoring (RQ 3A and 3B)

	Absolute monitoring accuracy after self-scoring			Absolute regulation accuracy after self-scoring		
	Multiplication, n = 581	Division, n = 386	Multiplication, n = 581	Division, n = 357	Multiplication, n = 581	Division, n = 357
	M0: intercept only	M0: intercept only	M0: intercept only	M0: intercept only	M1: effect of performance only	M1: effect of performance only
	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)
Fixed effects						
Intercept	0.10 (0.01)***	0.07 (0.02)***	0.06 (0.02)**	0.35 (0.02)***	0.47 (0.03)***	0.18 (0.02)***
Low/high <sup>a</sup>	-0.07 (0.03)*	0.02 (0.03)	0.02 (0.03)	-0.23 (0.04)***	-0.02 (0.04)	
Random effects						
$\sigma_e^2$ (day)	0.11 (0.02)***	0.06 (0.01)***	0.06 (0.01)***	0.31 (0.03)***	0.30 (0.03)***	0.11 (0.01)***
$\sigma_{u0}^2$ (student)	0.00 (0.01) <sup>b</sup>	0.01 (0.01)	0.01 (0.01)	0.01 (0.02)	0.01 (0.02)	0.04 (0.02)** <sup>d</sup>

Note: <sup>a</sup>0 = low-performing students; 1 = high-performing students; Unrounded values: <sup>b</sup> $\sigma_{u0}^2 = 0.003$ , <sup>c</sup> $\sigma_{u0}^2 = 0.002$ ; Effects with outliers: <sup>d</sup> $\sigma_e^2 = 0.21$  (0.03)\*\*\*  
 $\sigma_{u0}^2 = 0.04$  (0.02); <sup>e</sup> $\sigma_e^2 = 0.21$  (0.03)\*\*\*,  $\sigma_{u0}^2 = 0.04$  (0.02)  
 \*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

**Table S2.8**

*Effect of Monitoring Judgments on Regulation Judgments, Within the Whole Sample (RQ 4A and 4B)*

	<i>B</i> ( <i>SE</i> )	95% CI for Odds Ratio			<i>B</i> ( <i>SE</i> )	95% CI for Odds Ratio		
		Lower	Odds ratio	Upper		Lower	Odds ratio	Upper
<b>Before self-scoring</b>								
<b>Multiplication (n = 976)</b>								
Practice vs. Nothing	Intercept	3.95 (0.39)***			4.79 (0.59)***			
	Monitoring judgment	-0.99 (0.08)***	0.32	0.37	0.44	-1.18 (0.12)***	0.25	0.31
Instruction vs. Nothing	Intercept	4.65 (0.39)***			6.39 (0.65)***			
	Monitoring judgment	-1.40 (0.10)***	0.21	0.25	0.30	-1.67 (0.14)***	0.14	0.19
<b>After self-scoring</b>								
<b>Multiplication (n = 947)</b>								
Practice vs. Nothing	Intercept	4.56 (0.45)***			4.54 (0.45)***			
	Monitoring judgment	-1.19 (0.11)***	0.25	0.30	0.37	-1.20 (0.09)***	0.25	0.30
Instruction vs. Nothing	Intercept	5.38 (0.45)***			6.07 (0.56)***			
	Monitoring judgment	-1.66 (0.11)***	0.15	0.19	0.24	-1.70 (0.13)***	0.14	0.18

Note: "Practice" refers to regulation judgment = additional practice needed. "Instruction" refers to regulation judgment = additional instruction (and practice) needed. "Nothing" refers to regulation judgment = no intervention needed. \*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

**Table S2.9**  
*Effect of Monitoring Judgments on Regulation Judgments, Within the Low-Performing Students (RQ 4A and 4B)*

	<i>B (SE)</i>	95% CI for Odds Ratio			<i>B (SE)</i>	95% CI for Odds Ratio		
		Lower	Odds ratio	Upper		Lower	Odds ratio	Upper
<b>Before self-scoring</b>								
		<b>Multiplication (n = 272)</b>				<b>Division (n = 198)</b>		
Practice vs. Nothing	4.00 (0.60)***				3.19 (0.89)***			
		0.31	0.41	0.52			0.29	0.47
Monitoring judgment	-0.90 (0.15)***				-0.76 (0.24)***			0.75
Instruction vs. Nothing	4.79 (0.63)***				4.77 (0.94)***			
		0.20	0.28	0.36			0.21	0.34
Monitoring judgment	-1.29 (0.17)***				-1.07 (0.26)***			0.57
<b>After self-scoring</b>								
		<b>Multiplication (n = 261)</b>				<b>Division (n = 178)</b>		
Practice vs. Nothing	3.39 (0.52)***				5.82 (1.41)***			
		0.33	0.45	0.62			0.05	0.14
Monitoring judgment	-0.79 (0.16)***				-1.97 (0.54)***			0.40
Instruction vs. Nothing	4.39 (0.51)***				7.34 (1.44)***			
		0.16	0.24	0.36			0.04	0.11
Monitoring judgment	-1.43 (0.20)***				-2.23 (0.55)***			0.32

Note. "Practice" refers to regulation judgment = additional practice needed. "Instruction" refers to regulation judgment = additional instruction (and practice) needed. "Nothing" refers to regulation judgment = no intervention needed.

\*\*\*,  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

**Table S2.10**  
*Effect of Monitoring Judgments on Regulation Judgments, Within the High-Performing Students (RQ 4A and 4B)*

	<i>B</i> ( <i>SE</i> )	95% CI for Odds Ratio		<i>B</i> ( <i>SE</i> )	95% CI for Odds Ratio				
		Lower	Upper		Lower	Upper			
<b>Before self-scoring</b>									
		<b>Multiplication (n = 313)</b>			<b>Division (n = 200)</b>				
Practice vs. Nothing	Intercept	3.75 (0.86)***			8.98 (1.48)***				
	Monitoring judgment	-1.01 (0.17)***	0.26	0.36	0.51	-1.99 (0.29)***	0.08	0.14	0.24
Instruction vs. Nothing	Intercept	3.84 (0.92)**			- <sup>a</sup>				
	Monitoring judgment	-1.36 (0.19)***	0.18	0.26	0.37	-	-	-	-
<b>After self-scoring</b>									
		<b>Multiplication (n = 308)</b>			<b>Division (n = 191)</b>				
Practice vs. Nothing	Intercept	5.94 (1.48)***			9.108 (1.60)***				
	Monitoring judgment	-1.49 (0.31)***	0.12	0.23	0.42	-2.18 (0.32)***	0.06	0.11	0.21
Instruction vs. Nothing	Intercept	1.31 (0.76)			- <sup>a</sup>				
	Monitoring judgment	-1.05 (0.11)***	0.29	0.35	0.43	-	-	-	-

*Note:* "Practice" refers to regulation judgment = additional practice needed. "Instruction" refers to regulation judgment = additional instruction (and practice) needed. "Nothing" refers to regulation judgment = no intervention needed. <sup>a</sup> When outliers are excluded, no students within this sample made the "Instruction" judgment, thus this model could not be analyzed. Effects with outliers, division before self-scoring:  $B_{Intercept} = 7.58 (1.89)***$ ,  $B_{Monitoring\ judgment} = -2.25 (0.45)***$ , Odds Ratio = 0.11, 95% CI [0.04, 0.25]. Effects with outliers, division after self-scoring:  $B_{Intercept} = -1.74 (3.72)$ ,  $B_{Monitoring\ judgment} = -0.42 (0.69)$ , Odds Ratio = 0.66, 95% CI [0.17, 2.53].  
 \*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

**Table S2.11**  
*Effect of Performance (Continuous Variable) on Absolute Monitoring/Regulation Accuracy Before Self-Scoring*

	Absolute monitoring accuracy before self-scoring			Absolute regulation accuracy before self-scoring								
	Multiplication, <i>n</i> = 985			Division, <i>n</i> = 705			Multiplication, <i>n</i> = 973			Division, <i>n</i> = 692		
	M0: intercept only	M1: effect of performance	<i>B</i> ( <i>SE</i> )	M0: intercept only	M1: effect of performance	<i>B</i> ( <i>SE</i> )	M0: intercept only	M1: effect of performance	<i>B</i> ( <i>SE</i> )	M0: intercept only	M1: effect of performance	<i>B</i> ( <i>SE</i> )
Fixed effects												
Intercept	1.22 (0.04)***	2.07 (0.14)***	2.07 (0.14)***	0.96 (0.05)***	1.20 (0.11)***	1.20 (0.11)***	0.61 (0.02)***	0.83 (0.06)***	0.83 (0.06)***	0.43 (0.03)***	0.46 (0.06)***	0.46 (0.06)***
Performance		-0.23 (0.03)***	-0.23 (0.03)***		-0.08 (0.03)**	-0.08 (0.03)**		-0.06 (0.02)***	-0.06 (0.02)***		-0.01 (0.02)	-0.01 (0.02)
Random effects												
$\sigma^2_e$ (day)	1.25 (0.11)***	1.20 (0.11)***	1.20 (0.11)***	0.75 (0.07)***	0.75 (0.07)***	0.75 (0.07)***	0.36 (0.02)***	0.36 (0.02)***	0.36 (0.02)***	0.36 (0.04)***	0.36 (0.04)***	0.36 (0.04)***
$\sigma^2_{u0}$ (student)	0.10 (0.06)	0.03 (0.06)	0.03 (0.06)	0.14 (0.05)**	0.12 (0.05)**	0.12 (0.05)**	0.13 (0.02)***	0.12 (0.02)***	0.12 (0.02)***	0.08 (0.03)*	0.08 (0.03)*	0.08 (0.03)*

\*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

**Table S2.12**

*Effect of Performance (Continuous Variable) on Absolute Monitoring/Regulation Accuracy After Self-Scoring*

	Absolute monitoring accuracy after self-scoring			Absolute regulation accuracy after self-scoring								
	Multiplication, <i>n</i> = 954			Division, <i>n</i> = 676			Multiplication, <i>n</i> = 954			Division, <i>n</i> = 670		
	M0: intercept only	M1: effect of performance	M0: intercept only	M1: effect of performance	M0: intercept only	M1: effect of performance	M0: intercept only	M1: effect of performance	M0: intercept only	M1: effect of performance	M0: intercept only	M1: effect of performance
	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )
Fixed effects												
Intercept	0.12 (0.01)**	0.18 (0.03)***	0.09 (0.01)***	0.07 (0.02)**	0.44 (0.02)***	0.60 (0.05)***	0.36 (0.03)***	0.38 (0.05)***				
Performance		-0.02 (0.01) <sup>b</sup>		0.01 (0.01)		-0.04 (0.01)**		-0.01 (0.01)				
Random effects												
$\sigma_e^2$ (day)	0.13 (0.02)***	0.13 (0.02)***	0.06 (0.01)*** <sup>d</sup>	0.06 (0.01)*** <sup>d</sup>	0.34 (0.02)***	0.34 (0.02)***	0.32 (0.03)***	0.32 (0.03)***				
$\sigma_{u0}^2$ (student)	0.00 (0.01) <sup>a</sup>	0.00 (0.01) <sup>c</sup>	0.02 (0.01)* <sup>d</sup>	0.02 (0.01)* <sup>d</sup>	0.04 (0.03)	0.04 (0.03)	0.04 (0.03)	0.04 (0.02)				

Note: Unrounded values: <sup>a</sup>  $\sigma_{u0}^2 = 0.004$ , <sup>c</sup>  $\sigma_{u0}^2 = 0.003$ ; Effects with outliers: <sup>b</sup>  $B = -0.04$  (0.01)\*\*\*, <sup>d</sup>  $\sigma_e^2 = 0.15$  (0.05)\*\*\*,  $\sigma_{u0}^2 = 0.08$  (0.05);

\*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$



**Table S2.13**

*Mean Number of Omission and Commission Errors per Task, Including Standard Deviations Within Brackets*

	<b>Omission errors</b>	<b>Commission errors</b>
Multiplication		
Whole sample	0.62 (1.24)	1.72 (1.49)
Low-performing students	1.68 (1.71)	2.72 (1.73)
High-performing students	0.05 (0.26)	0.69 (0.64)
Division		
Whole sample	1.29 (1.69)	1.51 (1.55)
Low-performing students	2.69 (1.84)	2.39 (1.83)
High-performing students	0.06 (0.27)	0.64 (0.63)

**Table S2.14**  
Types of Needs in Percentages

	Additional instruction (and practice afterwards)	Additional practice		Additional instruction (and practice afterwards) or additional practice	No need for additional instruction or practice
		Because this student works too slowly	Because this student makes computational errors		
Multiplication					
Whole sample	32.2	5.2	9.7	17.9	35.0
Low-performing students	66.9	0.7	4.3	27.0	1.1 <sup>a</sup>
High-performing students	2.2 <sup>b</sup>	8.5	6.3	3.5	79.4
Division					
Whole sample	49.7	2.6	5.5	11.5	30.7
Low-performing students	88.6	-	-	11.4	-
High-performing students	2.6 <sup>b</sup>	6.6	6.1	4.1	80.6

<sup>a</sup>Sometimes, students who performed on average low, did not need additional practice or instruction on one of both days.

<sup>b</sup>Sometimes, student who performed on average high needed additional instruction on one of both days.

## Supplementary Materials - Chapter 3

### 3.1. Differences Between Analyses at the Within- and Between-student Level

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

Conclusions about whether or not students show accuracy awareness could be similar but can also differ depending on whether the analyses are conducted at the within or between-student level. Consider the theoretical example of student A and B displayed in Table S3.1. When analyzing this data at the between-student level, only including scores on the task on day 1, one would conclude that the students show *no* awareness of their judgment (in)accuracy, as the student of whom the judgment is more accurate (i.e., Student A) does not feel more confident about the accuracy of their judgment (i.e., both students rate their confidence as five). In contrast, when analyzing this data at the within-student level, one would conclude that both students *do* show accuracy awareness as they both feel more confident of their more accurate judgments. Hence, in the Results section, we presented the results of the analyses at the within-student level (i.e., day level). To enable comparison with the study by Nederhand et al. (2021), the results at the between-student level are reported below in Table S3.2.

**Table S3.1***Numerical Example of Judgment Accuracy and SOJs for Two Fictional Students*

	Student A		Student B	
	Accuracy	SOJ	Accuracy	SOJ
Day 1	0	5	1	5
Day 2	2	3	3	3

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

**Table S3.2***Effects of Absolute Monitoring/Regulation Accuracy on SOJ-m/SOJ-r Before Self-Scoring, at the Student Level*

	Whole sample		Low-performing students		High-performing students	
	B (SE)	R <sup>2</sup>	B (SE)	R <sup>2</sup>	B (SE)	R <sup>2</sup>
Monitoring						
Multiplication	0.01 (0.04)	.00	0.11 (0.06) <sup>a</sup>	.04	-0.36 (0.09) <sup>***</sup>	.10
Division	-0.14 (0.06) <sup>*</sup>	.02	-0.23 (0.10) <sup>*</sup>	.03	0.04 (0.12)	.00
Regulation						
Multiplication	-0.04 (0.06)	.00	-0.13 (0.11)	.01	-0.17 (0.10)	.01
Division	-0.17 (0.08) <sup>*</sup>	.01	-0.54 (0.15) <sup>***</sup>	.08	-0.28 (0.18) <sup>b</sup>	.03

Note. Some effects were significant when outliers were still included: <sup>a</sup>B = 0.14 (0.06),  $p = .011$ ;

<sup>b</sup>B = -0.36 (0.17),  $p = .038$

\*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

### Interpretation of the Results Displayed in Table S3.2

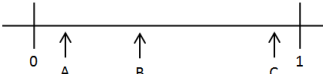
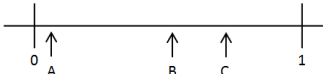




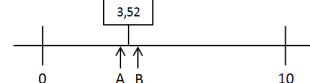

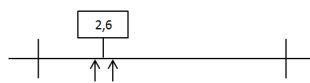

Table S3.2 shows the results of the analyses in which students' SOJ-m/SOJ-r were regressed on their absolute monitoring/regulation accuracy at the between-student level—that is, only measures of students' accuracy and SOJs of the first day were included. All significant effects at the between-student level were negative, as was the case for the significant effects at the within-student level (i.e., including data of both days, Table 3.2 in the Results section in Chapter 3). Negative significant effects at the between-student level mean that students who made a more accurate monitoring or regulation judgment on a task, felt more confident about the accuracy of their judgment, than students who made a less accurate judgment on that task (which is in our view no indication of accuracy awareness).

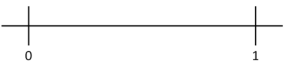

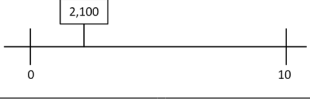
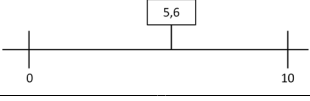
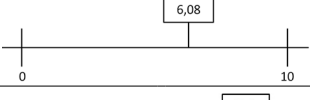

With regard to monitoring, three of the six effects were significant at the between-student level (Table S3.2), whereas only one of the six effects was significant when analyzing data at the within-student level (Table 3.2 in the Results section in Chapter 3). With regard to regulation, two of the six effects were significant at the between-student level, whereas three of the six effects were significant when analyzing data at the within-student level.

Remarkably, eight of the twelve regression coefficients were larger at the between-student level compared to the within-student level. For high-performing students' monitoring on the multiplication task the effect size in terms of  $R^2$  was considerably larger at the between-student level (i.e., .10; medium) than at the within-student level (i.e., .01; small; in terms of  $R^2$ , .01 is the criterion for a small effect, .09 for a medium effect, .25 for a large effect; Cohen, 1988). If we would have drawn conclusions about students' accuracy awareness based on the results at the between-student level, we could have unrightly concluded that the data showed clear signs of that specific subsets of students were (somewhat) aware of their monitoring and regulation accuracy.

## Supplementary Materials - Chapter 4

### 4.1. Test Items (Translated From Dutch) Including Misconception Indications

First assignment		Indication of misconception by erroneous answer options
Item		
1: Where do the numbers go on the number line? Circle the right answer, so circle A, B, or C.		
1a. Where does 0.9 go?		A/B = Whole number.
1b. Where does 0.07 go?		B/C = Ignoring zero in tenths place.
1c. Where does 0.756 go?		A/B = Fraction.
1d. Where does 0.13 go?		A/C = Outside 0 and 1.
1e. Where does 0.200 go?		B/C = Zero at end makes bigger.
2. Circle the right answer, so circle A or B. The number line now goes from 0 to 10!		
2a. The sign is at 4.3. Where does 4.482 go?		A = Fraction.
2b. The sign is at 3.52. Where does 3.8 go?		A = Whole number.
2c. The sign is at 7.2. Where does 7.05 go?		B = Ignoring zero in tenths place.
2d. The sign is at 2.6. Where does 2.400 go?		B = Zero at end makes bigger.
3. Mark the right place at the number line. The number line goes from 0 to 1.		
3a. Mark where 0.06 goes.		Placing 0.06 outside 0 and 1 = Outside 0 and 1. Placing it around 0.6 = Ignoring zero in tenths place

3b. Mark where 0.281 goes.		Placing 0.281 outside 0 and 1 = outside 0 and 1. Placing it at the right half of the number line = Zero at end makes bigger. Placing it close to zero = Fraction
3c. Mark where 0.400 goes.		Placing 0.400 outside 0 and 1 = outside 0 and 1. Placing it at the right half of the number line = Zero at end makes bigger. Placing it close to zero = Fraction.
4. Mark the right place at the number line. The number line goes from 0 to 10!		
4a. The sign is at 2.100. Mark where 2.7 goes.		Placing 2.7 left of the sign = Zero at end makes bigger.
4b. The sign is at 5.6. Mark where 5.39 goes.		Placing 5.39 right of the sign = Whole number.
4c. The sign is at 6.08. Mark where 6.3 goes.		Placing 6.3 left of the sign = Ignoring zero in tenths place.
4d. The sign is at 8.4. Mark where 8.739 goes.		Placing 8.739 left of the sign = Fraction.

## Second assignment

Item	Indication of misconception by erroneous answer options	
1: Where do the numbers go on the number line? Circle the right answer, so circle A, B, or C.		
1a. Where does 0.4 go?		A = Whole number. C = Fraction.
1b. Where does 0.21 go?		A/C = Outside 0 and 1.
1c. Where does 0.08 go?		B/C = Ignoring zero in tenths place.
1d. Where does 0.300 go?		B/C = Zero at end makes bigger.
1e. Where does 0.479 go?		A = Fraction. C = Whole number.
2. Circle the greatest number.		
2a. Circle the greatest number	0.47 0.3 0.926	$0.3 / 0.47 =$ Fraction.
2b. Circle the greatest number	0.87 0.9 0.695	$0.87 / 0.695 =$ Whole number.
2c. Circle the greatest number	0.001 0.1 0	0 = Outside 0 and 1.
2d. Circle the greatest number	0.300 0.40 0.6	$0.300 / 0.40 =$ Zero at end makes bigger.
2e. Circle the greatest number	0.2 0.09 0.007	$0.09 / 0.007 =$ Ignoring zero in tenths place.
3. Are the numbers smaller, equal or greater? Thick the right box.		



Item	Indication of misconception by erroneous answer options
3a. 0.70 is <input type="checkbox"/> smaller than <input type="checkbox"/> equal to 0.7 <input type="checkbox"/> greater than	Smaller than = Fraction. Greater than = Zero at end makes bigger.
3b. 0.54 is <input type="checkbox"/> smaller than <input type="checkbox"/> equal to 0.8 <input type="checkbox"/> greater than	Greater than = Whole number
3c. 0.03 <input type="checkbox"/> smaller than <input type="checkbox"/> equal to 0.3 <input type="checkbox"/> greater than	Equal to = Ignoring zero in tenths place.
3d. 0.4 is <input type="checkbox"/> smaller than <input type="checkbox"/> equal to 0.524 <input type="checkbox"/> greater than	Greater than = Fraction.
3e. 0.006 is <input type="checkbox"/> smaller than <input type="checkbox"/> equal to 0 <input type="checkbox"/> greater than	Smaller than = Outside 0 and 1
4. Write down a number.	
4a. Write down a number between 0.6 and 0.65.	Incorrect answers can indicate multiple misconceptions. These items were considered to test students' overall understanding of decimal place values; students were considered to only answer these items correctly if they had some basic understanding of the meaning of the different place values (i.e., that the first number after the decimal point reflects the tenths, the second number reflects the hundredths, etc.).
4b. Write down a number between 0.12 and 0.127.	

## 4.2. Coding Scheme for the Think-Aloud Transcripts

**Table S4.1**

*Coding Scheme for the Think-Aloud Transcripts*

<b>Codes per main category</b>	<b>Description</b>	<b>Example</b>
<b>Content</b>		
Item characteristics	Statements about characteristics or features of the items in the assignment(s), such as difficulty or physical appearance of the numbers, answer options, or problem type. N.B. Statements related to students' (mis)conceptions do not belong to this category.	... because of the pyramid form of the answers... ... this (item) is mean...
Curriculum	Statements about what was or was not yet taught in the curriculum thus far. N.B. Statements related to video instruction do not belong to this category.	Decimals are new to them. ... we taught this (milliliters) somewhat.
Instruction this lesson	Statements related to the extent to what students paid attention to or remembered the video instruction prior to making the assignments.	Well, she pays attention to that kind of videos. She did get that explanation.
<b>Student</b>		
General cognitive	Statements related to students' cognitive ability or skills, in general or not specifically related to mathematics, such as students' intelligence, language skills or learning disorders.	It is a clever girl. He has dyslexia.
Math general	Statements related to students' general math ability.	Actually, he is quite good in math. ... is one of my weak math students...
Other math domain	Statements related to students' skills in a specific mathematical domain, other than decimals, such as fractions, money, and geometry.	She is strong with fractions. There are some gaps in geometry, time, money. <i>Teacher is referring to a specific student.</i>
Effort and work regulation	Statements related to students' effort and regulation during working, such as speed, concentration, sloppiness and carefulness of working.	I think this has a lot to do with concentration. She is going to think really hard.

**Table S4.1***Continued*

<b>Codes per main category</b>	<b>Description</b>	<b>Example</b>
Affective	Statements related to students' emotions, motivation, and attitude, such as confidence, interest and stress.	... and so student x thinks... oh exciting! ... and student x likes it...
Class behavior	Statements related to students' general classroom behavior, not specifically related to working.	That one has ADHD.
Background	Statements related to students' background characteristics and home conditions, such as SES and characteristics of the parents.	... because he is an immigrant... ... or his parents do not speak much Dutch...
Gender	Statements related to a student's gender.	Well, I think, this is a he. Why do I assume this is a she?
Student other	Other statements about students that do not fall into one of the other categories. These are mostly very general statements.	That one is very unpredictable. Ok, that is a nice one. <i>Teacher is in both examples referring to a specific student.</i>
Fabricated student	Statements occurring in the answers-only condition implying that teachers had an idea about the identity of the student or tried to guess the identity.	I think somehow this is student x. I just immediately have a student in my mind.
<b>Answers</b>		
Item performance	Statements related to a student's performance on one item or a small group of items (max. five) in the first assignment, unrelated to students' strategy, understanding or (mis)conceptions.	He answered 1a correctly. It goes well half of the time. <i>Statement referring to student's performance on a subtask in the first assignment.</i>
Overall test performance	Statements related to the students' overall performance on the first assignment, unrelated to students' strategy, understanding or (mis)conceptions.	Here with practicing she is doing really bad. <i>The teacher is referring to the first assignment.</i> Well, her answers are so inconsequent.

**Table S4.1**

*Continued*

<b>Codes per main category</b>	<b>Description</b>	<b>Example</b>
<b>Student*Content</b> Understanding decimals	Statements related to students' prior knowledge or general understanding of decimal numbers.	No, this one does not really get how it works behind the decimal point. [... she chooses this one correctly], because this looks like what she knows...
Strategy	Statements related to the strategy or approach used by students, such as how to determine a position of a number on the number line or whether students use the strategy of adding digits to make two numbers equal of length.	... because she will puzzle on that number line like "four, oh that's less than five". I think he will just add a digit.
(Mis)conception	Statements related to the specific decimal magnitude (mis)conceptions students might have.	...locating 0.08 as 0.8... She sees 70 and 7 and thinks 70 is bigger than 7.
Student guessed	Statements reflecting that a teacher thinks a student guessed which answer is correctly, but that the student does not actually understand the content.	Maybe he guesses one correctly in this task... Or it is a coincidence (that the student made this item correctly)...
Comparison to other student	Statements referring to comparison of the student that is being judged to another student. N.B. We included this code in the Student*Content category, since teachers compared the students on characteristics related to the content, such as their understanding of decimals or misconceptions.	Student x also did that wrong... Well, when I assess this as correctly for the others [I should do it certainly for her].
<b>Teacher</b> Affective teacher	Statements reflecting teachers' affective experiences during the process, including statements about hope and astonishment.	... because I just hope she knows this... That is really frustrating. <i>Teacher is referring to her own emotions, not to those of a student.</i>
Meta process teacher	Statements related to teachers' meta thinking about the judgment process.	This is very hard. Or do I have to many high expectations?

**Table S4.1***Continued*

<b>Codes per main category</b>	<b>Description</b>	<b>Example</b>
Guessing	Statements reflecting that teachers do not know why they make certain judgments.	Well, I don't know why.... Then I am going to guess a bit...
<b>Miscellaneous</b> Judgment	Statements reflecting the mere prediction of a teacher about the students' correctness of a test item. N.B. When another code is also applicable to the segment that code is dominant and assigned instead of the judgment code.	She answers 2b incorrectly. ... so I think he will choose the right answer option here.
Other	This code is assigned when another code does not apply, but when it is clear what a statement means. For example, a teacher reads aloud an item or poses a question to the researcher.	Well, larger than, smaller than..... Let's have a look at these answers...
Unclear	This code is assigned when it is not clear what a teacher's statement refers to. This mostly applies to incomplete statements.	And then he will... Well, indeed you see...

## Supplementary Materials - Chapter 5

### 5.1. Explanations and Descriptive Statistics of Cue Measures

**Table S5.1.**

*Explanation of Student Cue Measures*

<b>Cue</b>	<b>Question (translated from Dutch)</b>	<b>Answer options (translated from Dutch)</b>	<b>Source on which question is based</b>
Conscientiousness	This student works conscientiousness during the normal mathematics lesson. <i>Examples: This student works orderly. This student works precisely.</i>	Strongly disagree to Strongly agree <sup>a</sup>	Big Five conscientiousness scale (Goldberg, 1992)
Effort	This student shows effort during normal mathematics lessons. <i>Examples: this student works hard; this student pays attention.</i>	Strongly disagree to Strongly agree <sup>a</sup>	Cf. Helwig et al. (2001)
General mathematics ability	This student is in general strong in mathematics	Strongly below average/ Below average/ Average/ Above average/ Strongly above average	Cf. Helwig et al. (2001)
Interest	This student is generally interested in mathematics.	Strongly disagree to Strongly agree <sup>a</sup>	Cf. Karing (2009)
Learning problems	Does this student have learning problems (no diagnosis needed)?	No learning problems/ Dyslexia/ Dyscalculia/ ADHD/ ADD/ Autism/ Language delay/ Other, namely...	Cf. Van de Pol et al. (2021)
Nationality	What is the country of Birth of this student/ the mother of this student/ the father of this student?	The Netherlands/ Another Country, namely...	Cf. Driessen et al. (2015), Van de Pol et al. (2021)

**Table S5.1.***Continued*

<b>Cue</b>	<b>Question (translated from Dutch)</b>	<b>Answer options (translated from Dutch)</b>	<b>Source on which question is based</b>
Self-concept	This student generally feels confident about their mathematical skills. <i>Examples: this student is convinced that he/she performs well on mathematical tasks and tests; this student knows that he/she can master the mathematics skills that he/she needs to learn.</i>	Strongly disagree to Strongly agree <sup>a</sup>	Perceived self-efficacy scale (Marsh et al., 2006)
Sex	Before the start of the experiment teachers were asked to provide the experimenter with a list of student names and their sex.	Open question.	

<sup>a</sup> Scale: Strongly disagree, disagree, agree, strongly agree

**Table S5.2.**  
Descriptive Statistics of the Cue Values for the Multiplication and Division Task

Cue	Range	Multiplication			Division		
		Total sample	Student cues only	Student+ performance cues	Total sample	Student cues only	Student+ performance cues
Performance cues							
Multiplication	0 to 6	3.36 (2.14)	3.27 (2.14)	3.45 (2.13)	3.35 (2.13)	3.26 (2.15)	3.46 (2.11)
Division	0 to 6	2.65 (2.38)	2.57 (2.37)	2.74 (2.39)	2.65 (2.37)	2.57 (2.36)	2.73 (2.39)
Student cues							
Conscientiousness	1 to 4	2.90 (0.79)	2.85 (0.79)	2.95 (0.80)	2.90 (0.80)	2.84 (0.79)	2.95 (0.81)
Effort	1 to 4	3.18 (0.67)	3.16 (0.67)	3.20 (0.66)	3.18 (0.67)	3.16 (0.67)	3.19 (0.67)
Sex	boy/girl <sup>a</sup>	49.6/50.4%	50.0/50.0%	49.0/51.0%	49.4/50.6%	50.8/49.2%	47.8/52.2%
Mathematics ability	1 to 5	3.27 (1.10)	3.21 (1.12)	3.32 (1.08)	3.26 (1.09)	3.21 (1.12)	3.30 (1.06)
Interest	1 to 4	3.00 (0.71)	2.96 (0.69)	3.03 (0.73)	2.99 (0.71)	2.96 (0.69)	3.02 (0.73)
Learning problems	no/yes	75.5/24.5%	74.0/26.0%	77.0/23.0%	75.0/25.0%	73.4/26.6%	76.5/23.5%
Self-concept	1 to 4	2.83 (0.79)	2.78 (0.80)	2.89 (0.77)	2.82 (0.79)	2.77 (0.80)	2.88 (0.78)
Nationality <sup>b</sup>	1 to 5	0.22 (0.73)	0.24 (0.77)	0.21 (0.70)	0.22 (0.73)	0.24 (0.77)	0.21 (0.69)

<sup>a</sup> This was an open question, but teachers only gave these two answers. <sup>b</sup> 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).



## 5.2. Supplementary Results

**Table S5.3.**

*Effect of Availability of Performance Cues on Teachers' Absolute Judgment Accuracy (RQ1)*

	Multiplication, <i>n</i> = 532		Division, <i>n</i> = 531	
	M0: intercept only	M1: effect of condition	M0: intercept only	M1: effect of condition
Fixed effects	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )
Intercept	1.37 (0.07)***	2.01 (0.23)***	1.43 (0.07)***	2.45 (0.21)***
Condition <sup>a</sup>		-0.43 (0.13)***		-0.69 (0.12)***
Random effects	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )
$\sigma^2_e$ (student)	1.73 (0.16)***	1.68 (0.15)***	1.70 (0.12)***	1.57 (0.11)***
$\sigma^2_{u0}$ (teacher)	0.04 (0.03)	0.04 (0.03)	0.04 (0.03)	0.05 (0.03)

<sup>a</sup> Coded as follows: 0 = student cues only; 1 = student+performance cues

\*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

**Table S5.4**

*Unstandardized Effects of Cues on Teachers' Judgment, as an Indication of Teachers' Cue Use (RQ2)*

	Multiplication						Division					
	Student cues only <i>n</i> = 235		Student+performance cues <i>n</i> = 222		Student cues only <i>n</i> = 242		Student+performance cues <i>n</i> = 225		Student cues only <i>n</i> = 242		Student+performance cues <i>n</i> = 225	
	M0: intercept only	M1: effect of cues	M0: intercept only	M1: effect of cues	M0: intercept only	M1: effect of cues	M0: intercept only	M1: effect of cues	M0: intercept only	M1: effect of cues	M0: intercept only	M1: effect of cues
Fixed effects	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )
Intercept	3.82 (0.18)***	-0.74 (0.61)	3.91 (0.17)***	-0.10 (0.36)	2.99 (0.20)***	-2.13 (0.47)***	2.98 (0.22)***	-0.76 (0.23)***	2.98 (0.22)***	-2.13 (0.47)***	2.98 (0.22)***	-0.76 (0.23)***
Prior mul. perf. <sup>a</sup>				0.68 (0.03)***				0.10 (0.03)***				0.10 (0.03)***
Prior division perf. <sup>b</sup>				0.06 (0.02)**				0.71 (0.03)***				0.71 (0.03)***
Conscientiousness		0.20 (0.10)		0.17 (0.08)*		0.20 (0.12)		0.10 (0.06)		0.20 (0.12)		0.10 (0.06)
Effort		0.11 (0.15)		-0.12 (0.08)		0.12 (0.12)		0.02 (0.08)		0.12 (0.12)		0.02 (0.08)
Gender (boy / girl)		-0.03 (0.13)		0.01 (0.08)		-0.03 (0.14)		0.06 (0.09)		-0.03 (0.14)		0.06 (0.09)
Math ability		0.82 (0.10)***		0.20 (0.06)***		0.80 (0.12)***		0.28 (0.07)***		0.80 (0.12)***		0.28 (0.07)***
Interest		0.16 (0.12)		0.11 (0.11)		0.30 (0.14)*		-0.09 (0.08)		0.30 (0.14)*		-0.09 (0.08)
Learning problems		-0.08 (0.11)		-0.02 (0.11)		-0.18 (0.13)		-0.20 (0.09)*		-0.18 (0.13)		-0.20 (0.09)*
Self-concept		0.23 (0.10)*		0.15 (0.09)		0.30 (0.12)**		0.15 (0.07)*		0.30 (0.12)**		0.15 (0.07)*
Nationality		0.03 (0.09)		-0.03 (0.04)		0.03 (0.09)		0.02 (0.07)		0.03 (0.09)		0.02 (0.07)
Random effects	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )
$\sigma^2_e$ (student)	1.90 (0.22)***	0.59 (0.06)***	3.24 (0.32)***	0.31 (0.04)***	2.23 (0.27)***	0.69 (0.08)***	3.84 (0.37)***	0.34 (0.04)***	2.23 (0.27)***	0.69 (0.08)***	3.84 (0.37)***	0.34 (0.04)***
$\sigma^2_{u0}$ (teacher)	0.83 (0.28)**	0.52 (0.14)***	0.45 (0.27)	0.15 (0.05)**	1.10 (0.27)***	1.04 (0.19)***	1.06 (0.44)*	0.10 (0.04)*	1.10 (0.27)***	1.04 (0.19)***	1.06 (0.44)*	0.10 (0.04)*

<sup>a</sup>Prior multiplication performance, <sup>b</sup>Prior division performance

\*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

**Table S5.5**

*Unstandardized Effects of Cues on Students' Performance, as an Indication of Cue Diagnosticity (RQ3)*

	Multiplication						Division					
	Student cues only <i>n</i> = 243		Student+performance cues <i>n</i> = 232		Student cues only <i>n</i> = 242		Student+performance cues <i>n</i> = 233		Student+performance cues <i>n</i> = 233		Student+performance cues <i>n</i> = 233	
	M0: intercept only	M1: effect of cues	M0: intercept only	M1: effect of cues	M0: intercept only	M1: effect of cues	M0: intercept only	M1: effect of cues	M0: intercept only	M1: effect of cues	M0: intercept only	M1: effect of cues
Fixed effects	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )
Intercept	3.79 (0.17)***	-1.47 (1.13)	3.83 (0.17)***	-0.58 (0.48)	2.97 (0.25)***	-2.90 (1.11)**	3.06 (0.21)***	0.36 (0.86)	0.10 (0.06)	0.73 (0.05)***	-0.05 (0.18)	0.00 (0.23)
Prior mul. perf. <sup>a</sup>				0.53 (0.06)***								
Prior division perf. <sup>b</sup>				0.07 (0.05)								
Conscientiousness		0.55 (0.21)**		0.48 (0.19)*								
Effort		0.20 (0.35)		-0.32 (0.18)								
Gender (boy / girl)		0.44 (0.26)		0.29 (0.22)								
Math ability		0.31 (0.15)*		0.10 (0.14)								
Interest		0.21 (0.24)		0.35 (0.22)								
Learning problems		-0.44 (0.37)		-0.11 (0.25)								
Self-concept		0.32 (0.21)		0.06 (0.15)								
Nationality		0.11 (0.22)		0.10 (0.11)								
Random effects	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )
$\sigma^2_e$ (student)	3.85 (0.30)***	2.91 (0.23)***	4.04 (0.30)***	2.01 (0.22)***	4.91 (0.39)***	3.37 (0.28)***	5.29 (0.36)***	2.05 (0.31)***	0.00 (0.08)	0.00 (0.08)	0.00 (0.08)	0.00 (0.08)
$\sigma^2_{u0}$ (teacher)	0.53 (0.27)*	0.46 (0.23)*	0.43 (0.31)	0.01 (0.08)	1.38 (0.42)***	1.29 (0.39)***	0.81 (0.40)*					

<sup>a</sup> Prior multiplication performance, <sup>b</sup> Prior division performance

\*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

**Table S5.6***Effect of Availability of Performance Cues on Teachers' Confidence in Their Judgment (RQ5A)*

	<b>Multiplication, n = 533</b>		<b>Division, n = 532</b>	
	<b>M0: intercept only</b>	<b>M1: effect of condition</b>	<b>M0: intercept only</b>	<b>M1: effect of condition</b>
Fixed effects	<i>B (SE)</i>	<i>B (SE)</i>	<i>B (SE)</i>	<i>B (SE)</i>
Intercept	4.50 (0.08)***	3.95 (0.19)***	4.37 (0.08)***	3.48 (0.23)***
Condition <sup>a</sup>		0.37 (0.10)***		0.59 (0.14)***
Random effects	<i>SS (SE)</i>	<i>SS (SE)</i>	<i>SS (SE)</i>	<i>SS (SE)</i>
$\sigma^2_e$ (student)	0.78 (0.06)***	0.74 (0.05)***	0.96 (0.10)***	0.87 (0.07)***
$\sigma^2_{u0}$ (teacher)	0.15 (0.04)***	0.15 (0.04)***	0.15 (0.05)**	0.15 (0.05)**

<sup>a</sup> Coded as follows: 0 = student cues only; 1 = student+performance cues\*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

**Table S5.7**  
*Effect of Teachers' Absolute Judgment Accuracy on the Confidence in Their Judgment (RQ5B First Part)*

	Multiplication						Division	
	Student cues only <i>n</i> = 268		Student+performance cues <i>n</i> = 262		Student cues only <i>n</i> = 276		Student+performance cues <i>n</i> = 256	
	M0: intercept only	M1: effect of cues	M0: intercept only	M1: effect of cues	M0: intercept only	M1: effect of cues	M0: intercept only	M1: effect of cues
Fixed effects	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )
Intercept	4.26 (0.10)***	4.42 (0.11)***	4.67 (0.08)***	4.83 (0.10)***	4.06 (0.11)***	4.30 (0.13)***	4.68 (0.10)***	5.00 (0.09)***
Accuracy		-0.10 (0.03)**		-0.13 (0.04)**		-0.14 (0.04)***		-0.30 (0.04)***
Random effects	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )
$\sigma^2_e$ (student)	0.64 (0.06)***	0.63 (0.06)***	0.80 (0.09)***	0.78 (0.10)***	0.82 (0.09)***	0.79 (0.09)***	0.76 (0.08)***	0.66 (0.07)***
$\sigma^2_{u0}$ (teacher)	0.24 (0.08)***	0.23 (0.08)**	0.12 (0.05)**	0.11 (0.05)*	0.27 (0.07)***	0.25 (0.06)***	0.18 (0.08)*	0.16 (0.08) *

\*\*\*  $p \leq .001$ , \*\*  $p \leq .01$ , \*  $p \leq .05$

**Table S5.8**

*Effect of Availability of Performance Cues and Teachers' Absolute Judgment Accuracy on the Confidence in Their Judgment (RQ5B Second Part)*

	Multiplication, <i>n</i> = 547				Division, <i>n</i> = 513				
	M0: intercept only	M1: effect of accuracy and condition	M2: interaction effect	M0: intercept only	M1: effect of accuracy and condition	M2: interaction effect	M0: intercept only	M1: effect of accuracy and condition	M2: interaction effect
Fixed effects	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )	<i>B</i> ( <i>SE</i> )
Intercept	4.45 (0.08)***	4.21 (0.17)***	4.02 (0.20)***	4.36 (0.08)***	3.90 (0.25)***	3.58 (0.27)***			
Accuracy		-0.12 (0.03)***	0.02 (0.07)		-0.19 (0.03)***	0.04 (0.08)			
Condition <sup>a</sup>		0.27 (0.09)**	0.40 (0.12)***		0.48 (0.14)***	0.70 (0.15)***			
Accuracy*Condition			-0.10 (0.50)*			-0.16 (0.05)**			
Random effects	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )	<i>SS</i> ( <i>SE</i> )			
$\sigma^2_e$ (student)	0.80 (0.06)***	0.75 (0.05)***	0.75 (0.05)***	0.97 (0.10)***	0.82 (0.07)***	0.81 (0.07)***			
$\sigma^2_{u0}$ (teacher)	0.18 (0.05)***	0.18 (0.06)**	0.18 (0.06)**	0.15 (0.05)**	0.12 (0.05)**	0.12 (0.05)**			

<sup>a</sup> Coded as follows: 0 = student cues only; 1 = student+performance cues

\*\*\* *p* ≤ .001, \*\* *p* ≤ .01, \* *p* ≤ .05

## Supplementary Materials - Chapter 6

### 6.1. Explanation and Descriptive Statistics of Teachers' Perceptions of Student Characteristics

**Table S6.1.**

*Explanation and Descriptive Statistics of Teachers' Perceptions of Student Characteristics*

Student characteristic	Question (translated from Dutch)
Amount of Contact	I have a lot of contact with this student.
Conscientiousness	This student works conscientiousness during the normal mathematics lesson. <i>Examples: This student works orderly. This student works precisely.</i>
Effort	This student shows effort during normal mathematics lessons. <i>Examples: this student works hard; this student pays attention.</i>
Extraversion	This student is generally extravert in class. <i>Examples: this student is talkative; this student is <b>not</b> withdrawn.</i>
Interest	This student is generally interested in mathematics.
Mathematics ability	This student is in general strong in mathematics
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:...
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:...
Likeability	I like this student.
Self-concept	This student generally feels confident about their mathematical skills. <i>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.</i>
Sex	Before the start of the experiment teachers were asked to provide the experimenter with a list of student names and their sex.

<sup>a</sup> Scale: Strongly disagree, disagree, agree, strongly agree.

<sup>b</sup> Strongly below average, below average, average, above average, strongly above average

<sup>c</sup> 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).

<sup>d</sup> This was an open question, but teachers only gave these two answers

Source on which question is based	Range	Mean (SD)
-	1 to 4 <sup>a</sup>	2.96 (0.66)
Big Five conscientiousness scale (Goldberg, 1992)	1 to 4 <sup>a</sup>	2.93 (0.76)
Cf. Helwig et al. (2001)	1 to 4 <sup>a</sup>	3.19 (0.64)
Big Five extraversion scale (Goldberg, 1992)	1 to 4 <sup>a</sup>	2.83 (0.89)
Cf. Karing (2009)	1 to 4 <sup>a</sup>	3.04 (0.66)
Cf. Helwig et al. (2001)	1 to 5 <sup>b</sup>	3.39 (1.01)
Cf. Driessen et al. (2015) and Van de Pol et al. (2021)	1 to 5 <sup>c</sup>	0.24 (0.78)
Cf. Van de Pol et al. (2021)	no/yes	77.5/22.5%
-	1 to 5 <sup>b</sup>	3.72 (0.72)
Perceived self-efficacy scale (Marsh et al., 2006)	1 to 4 <sup>a</sup>	2.88 (0.76)
-	boy/girl <sup>d</sup>	53.1/46.9%

Sup



## References

- Ackerman, R. (2019). Heuristic cues for meta-reasoning judgments: Review and methodology. *Psihologjske Teme*, 28(1), 1-20. <https://doi.org/10.31820/pt.28.1.1>
- Adams, D. M., McLaren, B. M., Durkin, K., Mayer, R. E., Rittle-Johnson, B., Isotani, S., & Van Velsen, M. (2014). Computers in human behavior using erroneous examples to improve mathematics learning with a web-based tutoring system. *Computers in Human Behavior*, 36, 401-411. <https://doi.org/10.1016/j.chb.2014.03.053>
- Artelt, C., & Rausch, T. (2014). Accuracy of teacher judgments: When and for what reasons? In S. Krolak-Schwerdt, S. Glock, & M. Bohmer (Eds.), *Teachers' professional development: Assessment, training, and learning* (pp. 229-248). Sense Publishers.
- 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* ["Number & space junior" manual grade 6]. Noordhoff.
- Baars, M., Van Gog, T., De Bruin, A. B. H., & Paas, F. (2014a). 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/https://doi.org/10.1002/acp.3008>
- Baars, M., Vink, S., Van Gog, T., De Bruin, A., & Paas, F. (2014b). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learning and Instruction*, 33, 92-107. <http://dx.doi.org/10.1016/j.learninstruc.2014.04.004>
- Baars, M., Van Gog, T., De Bruin, A. B. H., & Paas, F. (2017). Effects of problem solving after worked example study on secondary school children's monitoring accuracy. *Educational Psychology*, 37(7), 810-834. <https://doi.org/10.1080/01443410.2016.1150419>
- Baars, M., van Gog, T., de Bruin, A. B. H., & Paas, F. (2018). Accuracy of primary school children's immediate and delayed judgments of learning about problem-solving tasks. *Studies in Educational Evaluation*, 58, 51-59. <https://doi.org/10.1016/j.stueduc.2018.05.010>
- Baars, M., Visser, S., Van Gog, T., De Bruin, A. B. H., & Paas, F. (2013). Completion of partially worked-out examples as a generation strategy for improving monitoring accuracy. *Contemporary Educational Psychology*, 38(4), 395-406. <https://doi.org/10.1016/j.cedpsych.2013.09.001>
- Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, 64, 417-444. <https://doi.org/10.1146/annurev-psych-113011-143823>
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5-31. <https://doi.org/10.1007/s11092-008-9068-5>
- Boekaerts, M., & Rozendaal, J. S. (2010). Using multiple calibration indices in order to capture the complex picture of what affects students' accuracy of feeling of confidence. *Learning and Instruction*, 20(5), 372-382. <https://doi.org/10.1016/j.learninstruc.2009.03.002>
- Borghouts, C., Buter, a., & Gool, A. (2019a). *Pluspunt 4 handleiding groep 6* ["Plus 4" manual grade 6]. Malmberg.
- Borghouts, C., Buter, a., & Gool, A. (2019b). *De wereld in getallen 5 handleiding groep 6* ["The world in numbers 5" manual grade 6]. Malmberg.
- Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62(3), 193-217. <https://doi.org/10.1037/h0047470>
- Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. University of California Press.

- Byers, J. L., & Evans, T. E. (1980). *Using a lens-model analysis to identify the factors in teacher judgment* (Research Series No. 73). Institute for Research on Teaching, Michigan State University. <https://eric.ed.gov/?id=ED189576>
- Calhoun, M. B., Emerson, R. W., Flores, M., & Houchins, D. E. (2007). Computational fluency performance profile of high school students with mathematics disabilities. *Remedial and Special Education, 28*(5), 292-303. <https://doi.org/10.1177/07419325070280050401>
- 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>
- Campbell, C., & Levin, B. (2009). Using data to support educational improvement. *Educational Assessment, Evaluation and Accountability, 21*(1), 47-65. <https://doi.org/10.1007/s11092-008-9063-x>
- 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.
- Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. <https://doi.org/10.1037/0033-2909.112.1.155>
- Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation, 13*(5), 401-434. <https://doi.org/10.1080/13803610701728311>
- Crisp, V. (2008). The validity of using verbal protocol analysis to investigate the processes involved in examination marking. *Research in Education, 79*(1), 1-12. <https://doi.org/10.7227/RIE.79.1>
- De Bruin, A. B. H., Kok, E. M., Lobbestaal, 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. H., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology, 109*(3), 294-310. <https://doi.org/10.1016/j.jecp.2011.02.005>
- 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>
- De Smul, M., Heirweg, S., Devos, G., & Van Keer, H. (2019). School and teacher determinants underlying teachers' implementation of self-regulated learning in primary education. *Research Papers in Education, 34*(6), 701-724. <https://doi.org/10.1080/02671522.2018.1536888>
- Destan, N., & Roebbers, C. M. (2015). What are the metacognitive costs of young children's overconfidence? *Metacognition and Learning, 10*(3). <https://doi.org/10.1007/s11409-014-9133-z>
- Deunk, M. I., Smale-Jacobse, A. E., de Boer, H., Doolaard, S., & Bosker, R. J. (2018). Effective differentiation practices: A systematic review and meta-analysis of studies on the cognitive effects of differentiation practices in primary education. *Educational Research Review, 24*, 31-54. <https://doi.org/10.1016/j.edurev.2018.02.002>

- Dignath, C. & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition & Learning*, 3, 231-264. <https://doi.org/10.1007/s11409-008-9029-x>
- Dignath, C., Büttner, G. & Langfeldt, H.-P. (2008). How can primary school students acquire self-regulated learning most efficiently? A meta-analysis on interventions that aim at fostering self-regulation. *Educational Research Review*, 3, 101-129. <https://doi.org/10.1016/j.edurev.2008.02.003>
- 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/educ.2020.585683>
- Dignath, C. & Veenman, M. V. J. (2021). The role of direct strategy instruction in promoting self-regulated learning—evidence from classroom observation studies. *Educational Psychology Review*, 33(2), 489-533. <https://doi.org/10.1007/s10648-020-09534-0>
- Driessen, G., Elshof, D., Mulder, L., & Roeleveld, J. (2015). *Cohortonderzoek Cool 5-18: Technisch rapport basisonderwijs, derde meting 2013/14* [Cohort study COOL5-18: Technical report primary education, third measurement 2013/14]. ITS.
- Dufresne, A., & Kobasigawa, A. (1989). Children's spontaneous allocation of study time: Differential and sufficient aspects. *Journal of Experimental Child Psychology*, 47(2), 274-296. [https://doi.org/10.1016/0022-0965\(89\)90033-7](https://doi.org/10.1016/0022-0965(89)90033-7)
- Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, 22(4), 271-280. <https://doi.org/10.1016/j.learninstruc.2011.08.003>
- Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language*, 52(4), 551-565. <https://doi.org/10.1016/j.jml.2005.01.011>
- Durkin, K., & Rittle-Johnson, B. (2012). The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learning and Instruction*, 22(3), 206-214. <https://doi.org/10.1016/j.learninstruc.2011.11.001>
- Durkin, K., & Rittle-Johnson, B. (2015). Diagnosing misconceptions: Revealing changing decimal fraction knowledge. *Learning and Instruction*, 37, 21-29. <https://doi.org/10.1016/j.learninstruc.2014.08.003>
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. MIT Press.
- Fiorella, L., & Mayer, R. E. (2015). Eight ways to promote generative learning. *Educational Psychology Review*, 28, 717-741. <https://doi.org/10.1007/s10648-015-9348-9>
- Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, 1(3), 288-299. <https://doi.org/10.1037/0096-1523.1.3.288>
- Fitzsimmons, C. J., Thompson, C. A., & Sidney, P. G. (2020). Confident or familiar? The role of familiarity ratings in adults' confidence judgments when estimating fraction magnitudes. *Metacognition and Learning*, 15, 215-231. <https://doi.org/10.1007/s11409-020-09225-9>
- 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>
- Fritzsche, E. S., Händel, M., & Kröner, S. (2018). What do second-order judgments tell us about low-performing students' metacognitive awareness? *Metacognition and Learning, 13*(2), 159-177. <https://doi.org/10.1007/s11409-018-9182-9>
- Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science, 21*(3), 177-182. <https://doi.org/10.1177/0963721412445309>
- 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>
- Gabriele, A. J., Joram, E., & Park, K. H. (2016). Elementary mathematics teachers' judgment accuracy and calibration accuracy: Do they predict students' mathematics achievement outcomes? *Learning and Instruction, 45*, 49-60. <http://dx.doi.org/10.1016/j.learninstruc.2016.06.008>
- García, T., Rodríguez, C., González-Castro, P., González-Pienda, J. A., & Torrance, M. (2016). Elementary students' metacognitive processes and post-performance calibration on mathematical problem-solving tasks. *Metacognition and Learning, 11*(2), 139-170. <https://doi.org/10.1007/s11409-015-9139-1>
- 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>
- Gortazar, L., de Lafuente, D. M., & Vega-Bayo, A. (2022). Comparing teacher and external assessments: Are boys, immigrants, and poorer students undergraded? *Teaching and Teacher Education, 115*, 103725. <https://doi.org/10.1016/j.tate.2022.103725>
- Griffin, T. D., Mieliński, M. K., & Wiley, J. (2019). Improving students' metacomprehension accuracy. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 619-646). Cambridge University Press.
- Griffin, T. D., Wiley, J., & Salas, C. R. (2013). Supporting effective self-regulated learning: the critical role of monitoring. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 19-34). Springer.
- Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition, 36*(1), 93-103. <https://doi.org/10.3758/MC.36.1.93>
- Hadjidemetriou, G. (2021). *A VR approach for modelling the assessment bias of primary school PE teachers and educating them about it* [Master's thesis, University of Twente]. <http://essay.utwente.nl/88168/>
- Händel, M., & Dresel, M. (2018). Confidence in performance judgment accuracy: The unskilled and unaware effect revisited. *Metacognition and Learning, 13*(3), 265-285. <https://doi.org/10.1007/s11409-018-9185-6>
- Händel, M., & Fritzsche, E. S. (2016). Unskilled but subjectively aware: Metacognitive monitoring ability and respective awareness in low-performing students. *Memory & Cognition, 44*(2), 229-241. <https://doi.org/10.3758/s13421-015-0552-0>
- Hartwig, M. K., & Dunlosky, J. (2017). Category learning judgments in the classroom: Can students judge how well they know course topics? *Contemporary Educational Psychology, 49*, 80-90. <https://doi.org/10.1016/j.cedpsych.2016.12.002>
- Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

- Hedeker, D., Gibbons, R., du Toit, M., & Cheng, Y. (2008). *Supermix: Mixed effects models*. Scientific Software International.
- Heirweg, S., De Smul, M., Merchie, E., Devos, G., & Van Keer, H. (2022). The long road from teacher professional development to student improvement: A school-wide professionalization on self-regulated learning in primary education. *Research Papers in Education, 37*(6), 929-953. <https://doi.org/10.1080/02671522.2021.1905703>
- 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.
- Holstein, K., McLaren, B. M., & Alevan, V. (2019). Co-designing a real-time classroom orchestration tool to support teacher-AI complementarity. *Journal of Learning Analytics, 6*(2), 27-52. <https://doi.org/10.18608/jla.2019.62.3>
- 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. <https://doi.org/10.1037/1045-3830.22.2.115>
- Isotani, S., Adams, D., Mayer, R. E., Durkin, K., Rittle-Johnson, B., & McLaren, B. M. (2011). Can erroneous examples help middle-school students learn decimals? In D. Kloss, C. Gillet, D. Crespo García, R.M. Wild, & F. Wolpers (Eds.), *Towards Ubiquitous Learning: Proceedings of the 6th European Conference on Technology Enhanced Learning, 2011* (pp. 181-195). Springer.
- 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?]. [Doctoral dissertation, Université Grenoble Alpes]. <https://www.theses.fr/2019GREAH006.pdf>
- Johnston, O., Wildy, H., & Shand, J. (2019). A decade of teacher expectations research 2008-2018: Historical foundations, new developments, and future pathways. *Australian Journal of Education, 63*(1), 44-73. <https://doi.org/10.1177/0004944118824420>
- Kaiser, J., Möller, J., Helm, F., & Kunter, M. (2015). Das schülerinventar: Welche Schülermerkmale die Leistungsurteile von Lehrkräften beeinflussen [The student inventory: how student characteristics bias teacher judgments]. *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>
- Kaiser, J., Südkamp, A., & Möller, J. (2017). The effects of student characteristics on teachers' judgment accuracy: Disentangling ethnicity, minority status, and achievement. *Journal of Educational Psychology, 109*(6), 871-888. <https://doi.org/10.1037/edu0000156>
- Karing, C. (2009). Diagnostische Kompetenz von Grundschul- und Gymnasiallehrkräften im Leistungsbereich und im Bereich Interessen [Diagnostic competence of elementary and secondary school teachers in the domains of competence and interests]. *Zeitschrift für Pädagogische Psychologie, 23*(34), 197-209. <https://doi.org/10.1024/1010-0652.23.34.197>
- Karing, C., Pfost, M., & Artelt, C. (2011). Hängt die diagnostische Kompetenz von Sekundarstufenlehrkräften mit der Entwicklung der Lesekompetenz und der mathematischen Kompetenz ihrer Schülerinnen und Schüler zusammen? [Is there a relationship between lower secondary school teacher judgment accuracy and the development of students' reading and mathematical competence?]. *Journal for Educational Research Online, 3*(2), 119-147.

- Karst, K., Dotzel, S., & Dickhäuser, O. (2018). Comparing global judgments and specific judgments of teachers about students' knowledge: Is the whole the sum of its parts? *Teaching and Teacher Education, 76*, 194-203. <https://doi.org/10.1016/j.tate.2018.01.013>
- Kaufmann, E. (2020). How accurately do teachers' judge students? Re-analysis of Hoge and Coladarci (1989) meta-analysis. *Contemporary Educational Psychology, 63*, 101902. <https://doi.org/10.1016/j.cedpsych.2020.101902>
- Kistner, S., Rakoczy, K., Otto, B., Dignath, C., Büttner, G., & Klieme, E. (2010). Promotion of self-regulated learning in classrooms: investigating frequency, quality, and consequences for student performance. *Metacognition and Learning, 5*(2), 157-171. <https://doi.org/10.1007/s11409-010-9055-3>
- Klug, J., Bruder, S., Kelava, A., Spiel, C., & Schmitz, B. (2013). Diagnostic competence of teachers: A process model that accounts for diagnosing learning behavior tested by means of a case scenario. *Teaching and Teacher Education, 30*(1), 38-46. <https://doi.org/10.1016/j.tate.2012.10.004>
- Klug, J., Bruder, S., & Schmitz, B. (2016). Which variables predict teachers diagnostic competence when diagnosing students' learning behavior at different stages of a teacher's career? *Teachers and Teaching, 22*(4), 461-484. <https://doi.org/10.1080/13540602.2015.1082729>
- Kolovou, D., Naumann, A., Hochweber, J., & Praetorius, A. K. (2021). Content-specificity of teachers' judgment accuracy regarding students' academic achievement. *Teaching and Teacher Education, 100*, 103298. <https://doi.org/10.1016/j.tate.2021.103298>
- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349-370. <https://doi.org/10.1037/0096-3445.126.4.349>
- Kostons, D., Van Gog, T., & Paas, F. (2012). Training self-assessment and task-selection skills: A cognitive approach to improving self-regulated learning. *Learning and Instruction, 22*(2), 121-132. <https://doi.org/10.1016/j.learninstruc.2011.08.004>
- Kostons, D., & de Koning, B. B. (2017). Does visualization affect monitoring accuracy, restudy choice, and comprehension scores of students in primary education? *Contemporary Educational Psychology, 51*, 1-10. <https://doi.org/10.1016/j.cedpsych.2017.05.001>
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 121-1134. <https://doi.org/10.1037/0022-3514.77.6.1121>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. <https://doi.org/10.2307/2529310>
- Lipko, A. R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, K., & Cook, D. (2009). Using standards to improve middle school students' accuracy at evaluating the quality of their recall. *Journal of Experimental Psychology: Applied, 15*(4), 307-318. <https://doi.org/10.1037/a0017599>
- Lipko-Speed, A. R. (2013). Can young children be more accurate predictors of their recall performance? *Journal of Experimental Child Psychology, 114*(2), 357-363. <https://doi.org/10.1016/j.jecp.2012.09.012>
- 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](https://doi.org/10.1207/s15327574ijt0604_1)
- McClelland, M. M., & Cameron, C. E. (2011). Self-regulation and academic achievement in elementary school children. *New Directions for Child and Adolescent Development, 133*, 29-44. <https://doi.org/https://doi.org/10.1002/cd.302>
- 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>

- Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin and Review*, 15(1), 174–179. <https://doi.org/10.3758/PBR.15.1.174>
- Metcalfe, J., & Finn, B. (2013). Metacognition and control of study choice in children. *Metacognition and Learning*, 8(1), 19–46. <https://doi.org/10.1007/s11409-013-9094-7>
- Molenaar, I., Horvers, A., Dijkstra, R., & Baker, R. S. (2020). Personalized visualizations to promote young learners' SRL: the learning path app. *ACM International Conference Proceeding Series*, 330–339. <https://doi.org/10.1145/3375462.3375465>
- Molenaar, I., & Knoop-van Campen, C. (2017). Teacher dashboards in practice: Usage and impact. In Lavoué, E., Drachler, H., Verbert, K., Broisin, J., & Pérez-Sanagustín, M. (Eds.), *Data Driven Approaches in Digital Education: 12th European Conference on Technology Enhanced Learning* (pp. 125-138). Springer. [https://doi.org/10.1007/978-3-319-66610-5\\_10](https://doi.org/10.1007/978-3-319-66610-5_10)
- Muthén, L. K., & Muthén B. O. (1998-2017). *Mplus user's guide, 8th Edition*. Muthén & Muthén.
- Nederhand, M., Tabbers, H., de Bruin, A., & Rikers, R. (2021). Metacognitive awareness as measured by second-order judgements among university and secondary school students. *Metacognition and Learning*, 16(1), 1-14. <https://doi.org/10.1007/s11409-020-09228-6>
- Nuutila, K., Tuominen, H., Tapola, A., Vainikainen, M. P., & Niemivirta, M. (2018). Consistency, longitudinal stability, and predictions of elementary school students' task interest, success expectancy, and performance in mathematics. *Learning and Instruction*, 56, 73-83. <https://doi.org/10.1016/j.learninstruc.2018.04.003>
- Oudman, S., Van de Pol, J., Bakker, A., Moerbeek, M., & Van Gog, T. (2018). Effects of different cue types on the accuracy of primary school teachers' judgments of students' mathematical understanding. *Teaching and Teacher Education*, 76, 214-226. <https://doi.org/10.1016/j.tate.2018.02.007>
- Oudman, S., Van de Pol, J., Janssen, E., & Van Gog, T. (2022a). *Students' Monitoring and Regulation Accuracy Awareness* [Manuscript submitted for publication]. Department of Education, Utrecht University.
- Oudman, S., Van de Pol, J., & Van Gog, T. (2022b). Effects of self-scoring their math problem solutions on primary school students' monitoring and regulation. *Metacognition and Learning*, 17(1), 213-239. <https://doi.org/10.1007/s11409-021-09281-9>
- Oudman, 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, 103982. <https://doi.org/10.1016/j.tate.2022.103982>
- Oudman, S., Van de Pol, J., Van Loon, M., & Van Gog, T. (2022c). *Primary School Teachers' Judgment Accuracy of their Students' Self-monitoring and Self-regulation Skills* [Manuscript submitted for publication]. Department of Education, Utrecht University.
- OECD(2022). *Trendshaping education 2022*. OECD Publishing. <https://doi.org/10.1787/6ae8771a-en>
- Ostermann, A., Leuders, T., & Nückles, M. (2017). Improving the judgment of task difficulties: Prospective teachers' diagnostic competence in the area of functions and graphs. *Journal of Mathematics Teacher Education*, 21(6), 579-605. <https://doi.org/10.1007/s10857-017-9369-z>
- Paleczek, L., Seifert, S., & Gasteiger-Klicpera, B. (2017). Influences on teachers' judgment accuracy of reading abilities on second and third grade students: A multilevel analysis. *Psychology in the Schools*, 54(3), 228–245. <https://doi.org/10.1002/pits.21993>
- 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>
- Parsons, S. A., Vaughn, M., Scales, R. Q., Gallagher, M. A., Parsons, A. W., Davis, S. G., ... Allen, M. (2018). Teachers' instructional adaptations: A research synthesis. *Review of Educational Research*, 88(2), 205-242. <https://doi.org/10.3102/0034654317743198>

- Patterson, M. L., Foster, J. L., & Bellmer, C. D. (2001). Another look at accuracy and confidence in social judgments. *Journal of Nonverbal Behavior, 25*(3), 207-219. <https://doi.org/10.1007/s10919-009-0072-3>
- 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>
- Pit-ten Cate, I. M., Krolak-Schwerdt, S., & Glock, S. (2016). Accuracy of teachers' tracking decisions: Short-and long-term effects of accountability. *European Journal of Psychology of Education, 31*(2), 225-243. <https://doi.org/10.1007/s10212-015-0259-4>
- Prediger, S. (2008). The relevance of didactic categories for analysing obstacles in conceptual change: Revisiting the case of multiplication of fractions. *Learning and Instruction, 18*(1), 3-17. <https://doi.org/10.1016/j.learninstruc.2006.08.001>
- 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>
- Raaijmakers, S. F., Baars, M., Schaap, L., Paas, F., Van Merriënboer, J., & Van Gog, T. (2018). Training self-regulated learning skills with video modeling examples: Do task-selection skills transfer? *Instructional Science, 46*(2), 273-290. <https://doi.org/10.1007/s11251-017-9434-0>
- Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology, 19*(4-5), 559-579. <https://doi.org/10.1080/09541440701326022>
- Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48*(2), 335-360. <https://doi.org/10.3102/0002831210374874>
- Rijksoverheid [Dutch national government] (n.d.). Hoe is de groep van mijn kind op de basisschool samengesteld? [How is my child's primary school class composed?]. <https://www.rijksoverheid.nl/onderwerpen/basisonderwijs/vraag-en-antwoord/hoe-zijn-de-groepen-in-het-basisonderwijs-bo-samengesteld>
- Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*(2), 346-362. <https://doi.org/10.1037/0022-0663.93.2.346>
- Roebers, C. M., Krebs, S. S., & Roderer, T. (2014). Metacognitive monitoring and control in elementary school children: Their interrelations and their role for test performance. *Learning and Individual Differences, 29*, 141-149. <https://doi.org/10.1016/j.lindif.2012.12.003>
- Roebers, C. M., Mayer, B., Steiner, M., Bayard, N. S., & van Loon, M. H. (2019). The role of children's metacognitive experiences for cue utilization and monitoring accuracy: A longitudinal study. *Developmental psychology, 55*(10), 2077. <https://doi.org/10.1037/dev0000776>
- Roll, I., Alevin, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction, 21*(2), 267-280. <https://doi.org/10.1016/j.learninstruc.2010.07.004>
- Rowan, E., Van de Pol, J., Janssen, E., & Van Gog, T. (2023). *How do I know what you know? Promoting teachers' judgement accuracy of their students' text comprehension* [Manuscript in preparation]. Department of Education, Utrecht University
- Rutherford, T. (2017). Within and between person associations of calibration and achievement. *Contemporary Educational Psychology, 49*, 226-237. <https://doi.org/10.1016/j.cedpsych.2017.03.001>
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119-144. <https://doi.org/10.1007/bf00117714>
- Schildkamp, K., Poortman, C., Luyten, H., & Ebbeler, J. (2017). Factors promoting and hindering data-based decision making in schools. *School Effectiveness and School Improvement, 28*(2), 242-258. <https://doi.org/10.1080/09243453.2016.1256901>



- 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>
- Schraw, G., Kuch, F., & Gutierrez, A. P. (2013). Measure for measure: Calibrating ten commonly used calibration scores. *Learning and Instruction*, 24, 48-57. <https://doi.org/10.1016/j.learninstruc.2012.08.007>
- Seegers, G., & Boekaerts, M. (1993). Task motivation and mathematics achievement in actual task situations. *Learning and Instruction*, 3(2), 133-150. [https://doi.org/10.1016/0959-4752\(93\)90012-O](https://doi.org/10.1016/0959-4752(93)90012-O)
- Serra, M. J., & DeMarree, K. G. (2016). Unskilled and unaware in the classroom: College students' desired grades predict their biased grade predictions. *Memory and Cognition*, 44(7), 1127-1137. <https://doi.org/10.3758/s13421-016-0624-9>
- Shavelson, R. J., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. *Review of Educational Research*, 51(4), 455-498. <https://doi.org/10.3102/00346543051004455>
- SLO [Dutch Foundation for Curriculum Development] (2021). TULE Rekenen/Wiskunde [National Mathematics Curriculum]. <https://www.slo.nl/thema/meer/tule/rekenen-wiskunde/>
- Snow, R. E. (1968). Brunswikian approaches to research on teaching. *American Educational Research Journal*, 5(4), 475-489. <https://doi.org/10.3102/00028312005004475>
- Spruce, R., & Bol, L. (2015). Teacher beliefs, knowledge, and practice of self-regulated learning. *Metacognition and Learning*, 10(2), 245-277. <https://doi.org/10.1007/s11409-014-9124-0>
- Stiggins, R. J., & Chappuis, J. (2006). What a difference a word makes: Assessment for learning rather than assessment of learning helps students succeed. *Journal of Staff Development*, 27, 10-14. <http://downloads.pearsonassessments.com/ati/downloads/What-a-difference-a-word-makes.pdf>
- Strijbos, J. W., Martens, R. L., Prins, F. J., & Jochems, W. M. (2006). Content analysis: What are they talking about? *Computers & Education*, 46(1), 29-48. <https://doi.org/10.1016/j.compedu.2005.04.002>
- Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, 104(3), 743-762. <https://doi.org/10.1037/a0027627>
- Thiede, K. W., Brendefur, J. L., Carney, M. B., Champion, J., Turner, L., Stewart, R., & Osguthorpe, R. D. (2018). Improving the accuracy of teachers' judgments of student learning. *Teaching and Teacher Education*, 76, 106-115. <https://doi.org/10.1016/j.tate.2018.08.004>
- Thiede, K. W., Brendefur, J. L., Osguthorpe, R. D., Carney, M. B., Bremner, A., Strother, S., ... Jesse, D. (2015). Can teachers accurately predict student performance? *Teaching and Teacher Education*, 49, 36-44. <https://doi.org/10.1016/j.tate.2015.01.012>
- Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes*, 47(4), 331-362. <https://doi.org/10.1080/01638530902959927>
- Thiede, K. W., Oswald, S., Brendefur, J. L., Carney, M. B., & Osguthorpe, R. D. (2019). Teachers' judgments of student learning of mathematics. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 678-695). Cambridge University Press. <https://doi.org/10.1017/9781108235631.027>
- Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4, 25-29. <https://doi.org/10.1037/h0071663>
- Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., ... & Reynolds, T. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: A review of literature. *Journal for the Education of the Gifted*, 27(2-3), 119-145. <https://doi-org/10.1177/016235320302700203>
- Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review*, 32, 100374. <https://doi.org/10.1016/j.edurev.2020.100374>

- Van Buuren, S. (2012). *Flexible imputation of missing data*. CRC press.
- Van de Pol, J., de Bruin, A., van Loon, M., & van Gog, T. (2017, August). *The effect of cue availability on students' and teachers' judgment accuracy* [Paper presentation]. 17th biennial conference of the European Association for Research on Learning and Instruction, Tampere, Finland.
- Van de Pol, J., De Bruin, A. B. H., Van Loon, M. H., & Van Gog, T. (2019). Students' and teachers' monitoring and regulation of students' text comprehension: Effects of comprehension cue availability. *Contemporary Educational Psychology, 56*, 236–249. <https://doi.org/10.1016/j.cedpsych.2019.02.001>
- Van de Pol, J. & Oudman, S. (2022). *Teachers' Judgment Accuracy of Students' Monitoring Skills: A Conceptual and Methodological Framework and Explorative Study* [Manuscript submitted for publication]. Department of Education, Utrecht University.
- 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*, 103482. <https://doi.org/10.1016/j.tate.2021.103482>
- Van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review, 22*(3), 271–296. <https://doi.org/10.1007/s10648-010-9127-6>
- Van de Pol, J., Volman, M., & Beishuizen, J. (2011). Patterns of contingent teaching in teacher–student interaction. *Learning and Instruction, 21*(1), 46–57. <https://doi.org/10.1016/j.learninstruc.2009.10.004>
- Van de Pol, J., Volman, M., Oort, F., & Beishuizen, J. (2014). Teacher scaffolding in small-group work: An intervention study. *Journal of the Learning Sciences, 23*(4), 600–650. <https://doi.org/10.1080/10508406.2013.805300>
- Van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. (2015). Integrating data-based decision making, assessment for learning and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy & Practice, 22*(3), 324–343. <https://doi.org/10.1080/0969594X.2014.999024>
- Vandevelde, S., Vandenbussche, L., & Van Keer, H. (2012). Stimulating self-regulated learning in primary education: Encouraging versus hampering factors for teachers. *Procedia-Social and Behavioral Sciences, 69*, 1562–1571. <https://doi.org/10.1016/j.sbspro.2012.12.099>
- Van Gog, T., Hoogerheide, V., & Van Harsel, M. (2020). The role of mental effort in fostering self-regulated learning with problem-solving tasks. *Educational Psychology Review, 32*, 1055–1072. <https://doi.org/10.1007/s10648-020-09544-y>
- Van Gog, T., Kester, L., & Paas, F. (2011). Effects of concurrent monitoring on cognitive load and performance as a function of task complexity. *Applied Cognitive Psychology, 25*(4), 584–587. <https://doi.org/10.1002/acp.1726>
- Van Loon, M. H., de Bruin, A. B., van Gog, T., & van Merriënboer, J. J. (2013). The effect of delayed-JOLs and sentence generation on children's monitoring accuracy and regulation of idiom study. *Metacognition and Learning, 8*(2), 173–191. <https://doi.org/10.1007/s11409-013-9100-0>
- Van Loon, M. H., De Bruin, A. B. H., Van Gog, T., Van Merriënboer, J. J. G., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica, 151*, 143–154. <https://doi.org/10.1016/j.actpsy.2014.06.007>
- Van Loon, M. H., & Roebbers, 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 Someren, M. V., Barnard, Y. F., & Sandberg, J. A. (1994). *The think aloud method: A practical approach to modelling cognitive processes*. Academic Press.

- Van Zanten, M., Van den Brom-Snijders, P., Van den Bergh, J., Meier, R., & Vrolijk, A. (2007). *Reken-wiskundedidactiek: Hele getallen* [Mathematics didactics: Whole numbers]. ThiemeMeulenhoff.
- Webb, M. B. (2015). *Exploring the correlation between teachers' mindset and judgment accuracy to reveal the cues behind teachers' expectations* [Doctoral dissertation, Boise State University]. <https://scholarworks.boisestate.edu/td/949/>
- Whitmer, S. P. (1982, March). *A descriptive multimethod study of teacher judgment during the marking process* [Paper presentation]. The annual meeting of the American Educational Research Association: The Many Publics of Education and Educational Research, New York City, NY.
- William, D. (2011). *Embedded formative assessment*. Solution Tree Press.
- 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.
- Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., & Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assessment in Education: Principles, Policy & Practice*, 28(3), 228-260. <https://doi.org/10.1080/0969594X.2021.1884042>
- Zhu, C. (2019). *Understanding the formation and improving the accuracy of teacher judgment* [Doctoral dissertation, Universität Passau]. <https://opus4.kobv.de/opus4-uni-passau/frontdoor/index/index/docId/738>
- Zhu, C., & Urhahne, D. (2018). The use of learner response systems in the classroom enhances teachers' judgment accuracy. *Learning and Instruction*, 58, 255-262. <https://doi.org/10.1016/j.learninstruc.2018.07.011>
- 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.



## Curriculum Vitae

Sophie Oudman was born in 1990 in Groningen, The Netherlands. She completed her secondary education in 2008 at the Praedinius Gymnasium in Groningen. In 2012, she was part of the first group of students to graduate from the ALPO, that is, the combination of the primary school teacher training program at the University of Applied Sciences Utrecht and the Bachelor Educational Sciences at Utrecht University. In 2015 she graduated (cum laude) from the Research Master Educational Sciences: Learning in Interaction at Utrecht University. In 2016, she became a PromoDoc: working part-time as a primary school teacher at the Dr. Bosschool in Utrecht and part-time on her PhD project at Utrecht University. Her PhD project resulted in this dissertation about how to improve monitoring and regulation accuracy of students, which is important for effective self-regulated learning, as well as of teachers, which is important for effective adaptive teaching.

# List of Publications

## Journal Articles (reverse chronological)

- Oudman, 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, 103982. <https://doi.org/10.1016/j.tate.2022.103982>
- Oudman, S.**, van de Pol, J., Janssen, E., & van Gog, T. (2022). *Students' monitoring and regulation accuracy awareness* [Manuscript submitted for publication]. Department of Education, Utrecht University.
- Oudman, S.**, van de Pol, J., van Loon, M., & van Gog, T. (2022). *Primary school teachers' judgment accuracy of their students' self-monitoring and self-regulation skills* [Manuscript submitted for publication]. Department of Education, Utrecht University.
- Van de Pol, J. & **Oudman, S.** (2022). *Teachers' judgment accuracy of students' monitoring Skills: A conceptual and methodological framework and explorative study* [Manuscript submitted for publication]. Department of Education, Utrecht University
- Oudman, S.**, Van de Pol, J., & Van Gog, T. (2021). Effects of self-scoring their math problem solutions on primary school students' monitoring and regulation. *Metacognition and Learning*, 17(1), 213-239. <https://doi.org/10.1007/s11409-021-09281-9>
- Hendrickx, M., Mainhard, T., **Oudman, S.**, Cillessen, A. H. N., & Brekelmans, M. (2019). Leerkrachtgedrag en de kwaliteit van sociale relaties in de klas: de leerkracht als sociale referent voor de sociale status van de leerling. *Kind en Adolescent*, 40, 67-88. <https://doi.org/10.1007/s12453-018-00195-z>
- Oudman, S.**, Van de Pol, J., Bakker, A., Moerbeek, M., & Van Gog, T. (2018). Effects of different cue types on the accuracy of primary school teachers' judgments of students' mathematical understanding. *Teaching and Teacher Education*, 76, 214-226. <https://doi.org/10.1016/j.tate.2018.02.007>
- Mainhard, M. T.\*, **Oudman, S.\***, Hornstra, T. E., Bosker, R. J., & Goetz, T. (2018). Student emotions in class: The relative importance of teachers and their interpersonal relations with students. *Learning and Instruction*, 53, 109-119. <https://doi.org/10.1016/j.learninstruc.2017.07.011> [\*Equal contribution]
- Hendrickx, M. M. H. G., Mainhard, M. T., **Oudman, S.**, Boor-Klip, H. J., & Brekelmans, J. M. G. (2017). teacher behavior and peer liking and disliking: The teacher as a social referent for peer status. *Journal of Educational Psychology*, 109(4), 546-558. <https://doi.org/10.1037/edu0000157>

## Conference Contributions as Presenting Author (reverse chronological)

- Oudman, S.**, van de Pol, J., van Loon, M., & van Gog, T (2022, December). *Primary school teachers' judgments of their students' monitoring and regulation skills* [Paper presentation]. EARLI SIG16 conference 2022, online
- Oudman, S.**, van de Pol, J., & van Gog, T (2021, August). *Effects of self-scoring math problem solutions on 4th grade students' monitoring and regulation* [Paper presentation]. EARLI conference 2021, online.
- Oudman, S.**, van de Pol, J., & van Gog, T (2021, August). *Primary school students' awareness of their monitoring and regulation accuracy* [Paper presentation]. JURE conference 2021, online.

- Oudman, S.,** van de Pol, J., & van Gog, T (2021, July). *Effect van zelf reketaken nakijken op de zelfbeoordeling en -regulatie van leerlingen in groep 6* [Paper presentation]. Onderwijs Research Dagen 2021, online.
- Oudman, S.,** van de Pol, J., & van Gog, T. (2021, May). *Primary school teachers' monitoring accuracy of students' math performance and awareness of their accuracy: Effects of cue availability* [Paper presentation]. EARLI SIG16 conference 2021, online.
- Oudman, S.,** van de Pol, J., & van Gog, T (2020, February). *Inschattingssaccuratesse van leerlingen en leerkrachten bij rekenen in groep 6* [Paper presentation]. PO Conferentie 2020, Utrecht, Nederlands.





In the ICO Dissertation Series dissertations are published of graduate students from faculties and institutes on educational research within the ICO Partner Universities: Eindhoven University of Technology, Leiden University, Maastricht University, Open University of the Netherlands, Radboud University Nijmegen, University of Amsterdam, University of Antwerp, University of Ghent, KU Leuven, Université Catholique de Louvain, University of Groningen, University of Twente, Utrecht University, Vrije Universiteit Amsterdam, and Wageningen University, and formerly Tilburg University (until 2002).

*List update 23 March, 2023 (the list will be updated every year in January)*

449. Janssen, E.M. (13-11-2020) Teaching Critical Thinking in Higher Education: Avoiding, Detecting, and Explaining Bias in Reasoning Utrecht: Utrecht University
450. Van den Broek, E.W.R. (09-10-2020) Language awareness in foreign language education. Exploring teachers' beliefs, practices and change processes Nijmegen: Radboud University Nijmegen
451. Kasch, J.A. (09-10-2020) Scaling the unscalable? Interaction and support in open online education Heerlen: Open University of the Netherlands
452. Otten, M. (30-10-2020) Algebraic reasoning in primary school: A balancing act Utrecht: Utrecht University
453. De Vrind, E. (25-11-2020) The SpeakTeach method, Towards self-regulated learning of speaking skills in foreign languages in secondary schools: an adaptive and practical approach Leiden: Leiden University
454. Tacoma, S.G. (15-11-2020) Automated intelligent feedback in university statistics education Utrecht: Utrecht University
455. Boonk, L.M. (04-12-2020) Exploring, measuring, and evaluating parental involvement in vocational education and training Heerlen: Open University of the Netherlands
456. Kickert, R. (04-12-2020) Raising the bar: Higher education students' sensitivity to the assessment policy Rotterdam: Erasmus University Rotterdam
457. Van der Wal, N.J. (09-12-2020) Developing Techno-mathematical Literacies in higher technical professional education Utrecht: Utrecht University
458. Vaessen, B.E. (08-01-2021) Students' perceptions of assessment and student learning in higher education courses. Eindhoven: Eindhoven University of Technology
459. Maureen, I.Y. (15-01-2021) Story time in early childhood education: designing storytelling activities to enhance (digital) literacy development. Enschede: University of Twente
460. Van Alten, D.C.D. (19-03-2021) Flipped learning in secondary education history classrooms: what are the effects and what is the role of self-regulated learning. Utrecht: Utrecht University
461. Gestsdóttir, S.M. (08-03-2021) Observing history teaching: historical thinking and reasoning in the upper secondary classroom. Amsterdam: University of Amsterdam
462. Chim, H.Q. (30-03-2021) Physical Activity Behavior and Learning in Higher Education. Maastricht: Maastricht University
463. Krijnen, E. (15-04-2021) Family literacy in context: Exploring the compatibility of a family literacy program with children's homes and schools. Rotterdam Erasmus University Rotterdam

464. Stolte, M. (07-05-2021) (In)attention for creativity: Unraveling the neural and cognitive aspects of (mathematical) creativity in children. Utrecht: Utrecht University
465. Rathé, S. (12-05-2021) Focusing on numbers – An investigation of the role of children's spontaneous focusing on Arabic number symbols in early mathematical development. Leuven: KU Leuven
466. Theelen, H. (12-05-2021) Looking around in the classroom. Developing preservice teachers' interpersonal competence with classroom simulations. Wageningen: Wageningen University
467. De Jong, L.A.H. (20-05-2021) Teacher professional learning and collaboration in secondary schools. Leiden: Leiden University
468. Sincer, I. (20-05-2021) Diverse Schools, Diverse Citizens? Teaching and learning citizenship in schools with varying student populations. Rotterdam: Erasmus University Rotterdam
469. Slijkhuis, E.G.J. (20-05-2021) Fostering active citizenship in young adulthood. Groningen: University of Groningen
470. Groothuisen-Vrancken, S.E.A. (02-06-2021) Quality and impact of practice-oriented educational research. Utrecht: Utrecht University
471. Hingstman, M. (07-06-2021) Supporting struggling students: prevention and early intervention with Success for All. Groningen: University of Groningen
472. Gerdes, J. (14-06-2021) All inclusive? Collaboration between teachers, parents and child support workers for inclusive education in prevocational schools. Amsterdam: Vrije Universiteit Amsterdam
473. Bai, H. (18-06-2021) Divergent thinking in young children. Utrecht: Utrecht University
474. Wijinker, W. (23-06-2021) The Unseen Potential of Film for Learning: Film's Interest Raising Mechanisms Explained in Secondary Science and Math. Utrecht: Utrecht University
475. Brummer, L. (24-09-2021). Unrooting the illusion of one-size-fits-all feedback in digital learning environments. Groningen: University of Groningen
476. Veldman, M.A. (01-07-21) Better together, social outcomes of cooperative learning in the first grades of primary education. Groningen: University of Groningen
477. Wang, J. (06-07-2021) Technology integration in education: Policy plans, teacher practices, and student outcomes. Leiden: Leiden University
478. Zhang, X. (06-07-2021) Teachers' teaching and learning motivation in China. Leiden: Leiden University
479. Poort, I.C. (02-09-2021) Prepared to engage? Factors that promote university students' engagement in intercultural group work. Groningen: University of Groningen
480. Guo, P. (07-09-2021) Online project-based higher education Student collaboration and outcomes. Leiden: Leiden University
481. Jin, X. (21-09-2021) Peer feedback in teacher professional development. Leiden: Leiden University
482. Atherley, E.N. (27-09-2021) Beyond the struggles: Using social-developmental lenses on the transition to clinical training. Maastricht: Maastricht University
483. Martens, S.E. (15-10-2021) Building student-staff partnerships in higher education. Maastricht: Maastricht University
484. Ovbiagbonhia, R. (08-11-2021) Learning to innovate: how to foster innovation competence in students of Built Environment at universities of applied science. Wageningen: Wageningen University
485. Van den Boom-Muilenburg, S.N. (11-11-2021) The role of school leadership in schools that sustainably work on school improvement with professional learning communities. Enschede: University of Twente

486. Sachishal, M.S.M. (11-11-2021) Science interest - Conceptualizing the construct and testing its predictive effects on current and future behavior. Amsterdam: University of Amsterdam
487. Meeuwissen, S.N.E. (12-11-2021) Team learning at work: Getting the best out of interdisciplinary teacher teams and leaders. Maastricht: Maastricht University
488. Keijzer-Groot, A.F.J.M. (18-11-2021) Vocational identity of at-risk youth – Tailoring to support career chances. Leiden: Leiden University
489. Wolthuis, F. (25-11-2021) Professional development in practice. Exploring how lesson study unfolds in schools through the lens of organizational routines. Groningen: University of Groningen
490. Akkermans-Rutgers, M. (06-12-2021) Raising readers. Stimulating home-based parental literacy involvement in the first, second, and third grade. Groningen: University of Groningen
491. Hui, L. (06-12-2021) Fostering Self-Regulated Learning: The Role of Perceived Mental Effort. Maastricht: Maastricht university
492. Jansen, D. (08-12-2021) Shadow education in the Netherlands: The position of shadow education in the educational landscape and students' school careers. Amsterdam: University of Amsterdam
493. Kamphorst, F. (15-12-2021) Introducing Special Relativity in Secondary Education. Utrecht: Utrecht University
494. Eshuis, E.H. (17-12-2021) Powering Up Collaboration and Knowledge Monitoring: Reflection-Based Support for 21st-Century Skills in Secondary Vocational Technical Education. Enschede: University of Twente
495. Abacioglu, C. S. (18-01-2022) Antecedents, Implications, and Professional Development of Teachers' Multiculturalism. Amsterdam: University of Amsterdam
496. Wong, J. (21-01-2022) Enhancing Self-Regulated Learning Through Instructional Support and Learning Analytics in Online Higher Education. Rotterdam: Erasmus University Rotterdam
497. Schophuizen, M.J.F. (18-02-2022) Educational innovation towards organizational development: the art of governing open and online education in Dutch higher education institutions. Heerlen: Open University of the Netherlands
498. Smeets, L.H. (18-03-2022) The auditor learning curve: professional development through learning from errors and coaching. Maastricht: Maastricht University
499. Beuken, J.A. (25-03-2022) Waves towards harmony – learning to collaborate in healthcare across borders. Maastricht: Maastricht University
500. Radovic, S. (25-03-2022) Instructional design according to the mARC model: Guidelines on how to stimulate more experiential learning in higher education. Heerlen: Open University of the Netherlands
501. Delnoij, L.E.C. (01-04-2022) Self-assessment for informed study decisions in higher education: A design-based validation approach. Heerlen: Open University of the Netherlands
502. Pieters, J. (01-04-2022) Let's talk about it: Palliative care education in undergraduate medical curricula. Maastricht: Maastricht University
503. Vulperhorst, J.P. (08-04-2022) Students' interest development prior, during and after the transition to a higher education programme. Utrecht: Utrecht University
504. Aben, J.E.J. (21-04-2022) Rectifying errors: A reconceptualization of the role of errors in peer-feedback provision and processing. Groningen: University of Groningen
505. Kooloos, C.C. (19-05-2022) Eye on Variety: The teacher's work in developing mathematical whole-class discussions. Nijmegen: Radboud University

506. El Majidi, A. (20-05-2022) Debate as a Tool for L2 Learning. Investigating the Potential of In-Class Debates for Second Language Learning and Argumentation Skills. Utrecht: Utrecht University
507. Van Halem, N. (24-05-2022) Accommodating Agency-Supportive Learning Environments in Formal Education. Amsterdam: Vrije Universiteit Amsterdam
508. Lee, Y.J. (8-06-2022) The medical pause in simulation training. Maastricht: Maastricht University
509. Loopers, J.H. (14-06-2022) Unravelling the dynamics of intrinsic motivation of students with and without special educational needs. Groningen: University of Groningen
510. Neroni, J. (17-06-2022) The Adult Learning Open University Determinants (ALoud) study: psychosocial factors predicting academic success in adult distance education. Heerlen: Open University of the Netherlands
511. Schlatter, E. (17-06-2022) Individual differences in children's scientific reasoning. Nijmegen: Radboud University
512. Bransen, D. (22-06-2022) Beyond the self: A network perspective on regulation of workplace learning. Maastricht: Maastricht University
513. Valero Haro, A. (27-06-2022) Fostering Argumentation with Online Learning Systems in Higher Education. Wageningen: Wageningen University
514. Schat, E. (05-07-2022) Integrating intercultural literary competence: An intervention study in foreign language education. Utrecht University
515. Soppe, K.F.B. (07-07-2022) To Match or not to Match? Improving Student-Program Fit in Dutch Higher Education. Utrecht: Utrecht University
516. Biver, F. (08-07-2022) Supporting students to study smart – a learning sciences perspective. Maastricht: Maastricht University
517. Weijers, R.J. (15-09-2022) Nudging Towards Autonomy: The effect of nudging on autonomous learning behavior in tertiary education. Rotterdam: Erasmus University Rotterdam
518. Van der Linden, S. (14-10-2022) Supporting teacher reflection in video-coaching settings. Enschede: University of Twente
519. Wijnen, F. (28-10-2022) Using new technology and stimulating students' higher-order thinking: a study on Primary school teachers' attitudes. Enschede: University of Twente
520. De Jong, W.A. (02-11-2022) Leading collaborative innovation in schools. Utrecht: Utrecht University
521. Bron, R. (04-11-2022) Collaborative course design in higher education – a team learning perspective. Enschede: University of Twente
522. De Vries, J.A. (18-11-2022) Supporting teachers in formative assessment in the classroom. Enschede: University of Twente
523. Le, T.I.N.H. (29-11-22) Towards a democratic school. Leiden: Leiden University
524. Wildeman, E. (30-11-2022) Vocational teachers' integrated language teaching. On the role of language awareness and related teaching behaviour. Eindhoven: Eindhoven University of Technology
525. Wolterinck, C.H.D. (2-12-2022), Teacher professional development in assessment for learning. Enschede: University of Twente
526. Kuang, X. (07-12-2022) Hypothesis generation, how to put students into motion. Enschede: University of Twente
527. Emhardt, S.N. (15-12-2022) You see? Investigating the effects of different types of guidance in eye movement modeling examples. Heerlen: Open University of the Netherlands



## Dankwoord

Toen ik, net als de kinderen in dit proefschrift, in groep 6 zat, wilde ik juf worden. Dat is gelukt! En ik wilde absoluut nooit promoveren, zo ongeveer tot het moment dat ik eraan begon. Maar wat heb ik ervan genoten! Jan van Tartwijk, bedankt dat je bleef vragen of ik toch niet mee wilde doen met 'een lollig projectje' genaamd PromoDoc (promoverend docent). Wat heb ik ontzettend veel geleerd de afgelopen jaren, dankzij een heleboel mensen.

Janneke en Tamara (ik twijfel tot het laatste moment of ik dit zal vervangen door 'Tammie en Jannie'), wat een fantastische begeleiders zijn jullie. Dank voor jullie bevoegenheid, alle tijd die jullie in mijn werk staken en de vrijheid die jullie me gaven. Ik heb bij jullie altijd de ruimte gevoeld om alles te kunnen vragen, om al m'n gedachten en emoties te uiten. Dat leidde regelmatig tot verhitte discussies, dat was smullen. Sorry dat ik soms wat doorsloeg in 'managing your supervisors' en dat de tranen, vooral in het begin, rijkelijk vloeiden (tijdens een meeting viel een keer: "Wat goed dat je het droog houdt"). Jullie wisten me ook altijd weer op te peppen als ik het nut van m'n eigen onderzoek niet meer zag. Janneke, jij leerde mij om niks voor waar aan te nemen tot het bewezen is en je gaf me vertrouwen in mijn eigen kunnen. Wat was het leuk om samen met jou te werken aan 'smeuïge' analyses. Tamara, jij had regelmatig een probleem alweer opgelost voordat ik het goed en wel uit de doeken had gedaan. Ook heb je mij geleerd dat als ik denk dat ik iets kort en bondig heb opgeschreven het altijd nóg korter kan.

Arthur, jij hoorde in de eerste twee jaar ook bij het begeleidingsteam en zowel jouw inhoudelijke begeleiding, onder andere bestaande uit veel kritische bevragingen en semantische discussies, als de emotionele begeleiding die je bood (d.m.v. uitspraken als "research is like walking in de fog") waren goud waard.

Anique, Paul, Liesbeth en Jan, bedankt voor het beoordelen van mijn proefschrift. Katharina, vielen Dank für die Bewertung meiner Dissertation.

Onmisbaar voor dit proefschrift waren de deelnemende leerkrachten en leerlingen, veel dank voor jullie tijd en moeite. En zonder onderzoeksassistenten hadden we nooit zoveel klassen kunnen bezoeken: Susan, Anna, Robbert, Jonne B, bedankt dat jullie afreisden naar de uithoeken van het land (tot plekken waar geen OV kwam en de conciërge jullie met de auto van een stationnetje kwam halen), transcribeerden, codeerden, scanden, en hielpen orde in de chaos te scheppen. Anne, bedankt voor je hulp bij het checken van de datapackages. Mariëtte, wat fijn dat je meewerkt aan een van de studies, zo enthousiast bent, en dat je tijdens conferenties van die slimme en grappige opmerkingen maakt.

In zeven jaar verslijt je heel wat kamergoten. David, Mei, Anne, Mare, Christa, Karin, Katrijn, Xiaojing, Jonne V, Marloes, Pierre, Dannie, Susan, Sophia, Michaela, Anouk, Yuanyuan, Angela, Linda, Jonne B, Rowan, Florence, Jane, Simone, Rik en Melis: wat een verademing om samen, asynchroon, min of meer hetzelfde proces te doorlopen, elkaar hierbij te steunen en veel te lachen. In onze kamer bleken we een hoop fysiek af te kunnen reageren: basketbalcompetities, buikspierkwartiertjes en het afschieten van stressraketten. Mede PromoDocs David, Anne, Mare, en Marloes, wat fijn dat we elkaar begrepen als we worstelden met de PhD-school-privé balans, waarbij school altijd voorrang kreeg. David, bedankt dat je me uitlachte als ik echt hele domme dingen zei. Sophia, waar zouden we geweest zijn zonder jouw lieve acties en de door jouw georganiseerde uitjes? En voor de nieuwe lichting kamergenoten: zonder deze boomer (ik voel me nog steeds beledigd) aan jullie zijde kunnen jullie het ook!

Andere (oud)collega's van de afdeling Educatie, dank dat jullie samen voor zo'n warm, inspirerend en leerzaam bad zorgen. Onder andere Margot, Martine, Luce, Bjorn, Larike, Tim, Caroline, Steven, Jael, Brechje en Vincent, dank voor de babbels, betrokkenheid, wijze adviezen en relativerende opmerkingen aan de koffietafel en tijdens feestjes. Eva J, bedankt voor de 387 Gutenberg bakkies, waarbij de koffie telkens weer over de rand klotste, vanwege ons wilde enthousiasme of gemekker. Ik bewonder jouw talent voor het stellen van de juiste vragen.

Collega's van de dr. Bosschool, dank voor de samenwerking en jullie harde werken voor de klas. Myra, bedankt dat je mij hebt aangenomen. Angelique, van jou heb ik het lesgeven geleerd. Jessika, jij leerde mij hoe om te gaan met de wat meer uitdagende interacties in de klas. Lieke, door jou werd mijn onderwijs innovatiever en creatiever. Danaë, bedankt voor het tekenen van de prachtige cover. Ik mis de vrijdagmiddagborrels met de collega's van de bovenbouw.

Naast collega's zijn er ook veel vrienden van wie ik heb mogen genieten ter afwisseling van het harde werken. Iedereen met wie ik hard heb gelachen, lekker heb gegeten, gesprongen, geklommen, gerend, gewandeld, gedroomd, of bij wie ik mijn hart heb mogen luchten: dankjewel! Een aantal vrienden wil ik in het bijzonder bedanken.

Lieve ALPO-vriendinnen Andrea, Janice, Marloes, Nette en Sietske, wat is het fijn om me zo verbonden met jullie te voelen, zo vaak gezellig met jullie te eten en over onderwijs te praten, zonder dat elke zin hoeft te worden afgesloten met een referentie. Minder belangrijk, maar jullie waren ook een grote hulp bij de totstandkoming van dit proefschrift, zoals door te helpen werven van leerkrachten, pilots draaien bij jullie in de klas, helpen coderen omwille van de interbeoordelaarsbetrouwbaarheid (sorry voor

het lange frustrerende proces Nette), lezen van stukken, of komen oppassen zodat ik kon werken.

Lieve Emma, Eva Z, Floor en Sanne, wat maken we al meer dan 20 jaar lang een hoop mee met elkaar. Altijd als we elkaar zien voelt dat als thuiskomen. Jessie, ik mis je. Eefje, al langer dan ik me kan herinneren ben jij mijn vriendin, bij jou heb ik aan een half woord genoeg.

Een bedankje aan alle speeltuinvrienden is ook wel op z'n plek, waar Heleen, Pim, Micha en Timo met kop en schouders bovenuit steken. Jullie helpen mij door de dagen waarop ik niet werk heen te slepen.

Eva V, elk advies dat jij geeft volg ik op. Hopelijk gaan er nog veel komen.

Lieve familie, ik voel me zo rijk met jullie. Papa en mama, van jullie leerde ik al vroeg dat hoge cijfers en diploma's niet belangrijk (zouden moeten) zijn, maar je best doen en talenten benutten wel. Mama, jij stimuleerde me de ALPO te gaan doen. Jij zag de leerkracht in mij en wat was het mooi geweest als je had geweten dat ik dat daadwerkelijk was geworden. Van jou heb ik ook de daadkracht meegekregen die maakt dat ik naast het lesgeven ook een gezin en proefschrift produceerde. Papa, dank voor alles, voor dat je altijd en onvoorwaardelijk voor me klaarstaat, voor je interesse en vertrouwen in mij, en dat je rustig blijft als ik dat niet ben. Margreet, wat fijn dat je er bent. Ik heb het gevoel altijd bij je terecht te kunnen als er wat is. Tho en Lau, broer en zus, wat een eer dat jullie mijn paranimfen willen zijn. Jullie inspireren mij allebei om in het nu te leven en goed voor mijn gezin te zorgen. Ik houd zielsveel van jullie, net als van de rest van jullie gezinnen: Hans Pieter, Guus, Twan en Cas, jullie moesten eens weten hoe vaak jullie worden geciteerd als opvoedkundige voorbeelden bij ons thuis. Rebekka, Rivka en Matthias, wat heerlijk om jullie zo vaak te zien en hopelijk gaan we nog vaak kamperen samen. Ook alle familie iets verder weg, ik heb maar geluk met jullie!

Tot slot mijn lieve thuis, mijn gezin. Tibbe en Abel, met jullie spelen geeft mij energie, jullie grapjes zijn het grappigst, jullie om me heen relateert de rest van mijn zorgen. Ik zal nooit genoeg krijgen van jullie eigenzinnigheid, fantasieën en snotkusjes. Het saaie boek is nu echt af! Liefste Robin, met jou dichtbij is het leven een feest. Dankjewel voor de bakken vol liefde die ik van jou krijg, dat je altijd zo relaxt én energiek bent, dat jij mij vaak beter kent dan ik mezelf, en dat het zo leuk blijft om samen met jouw op stenen te klimmen.

## Toetje

Lekker eten is fijn ter afwisseling van het harde werken. Bovendien hebben onze hersenen suikers nodig om te denken. Daarom, en om de kans te vergroten dat dit boekje nog eens wordt opengeslagen, deze toegift: een recept voor *Instant Chocolademousse*, lekker donker en stevig. Het recept is van Nigella Lawson en voor 6 tot 8 personen.

- 150 g marshmallows in stukjes gesneden (of minimmarshmallows)
- 50 g boter
- 250 g pure chocola ( $\geq$  70% cacao) in stukjes gehakt
- 60 ml kokend water
- 275 ml slagroom
- 1 tl vanille-extract

Doe de marshmallows, boter, chocola en water in pan met dikke bodem. Zet de pan op laag vuur en laat, terwijl je af en toe roert, alles smelten. Zet de pan van het vuur.

Klop de slagroom met vanille-extract stijf en spatel het door de chocola tot je een gladde samenhangende massa hebt.

Verdeel de mousse over kleine glaasjes en zet in de koelkast om af te laten koelen, maakt niet uit hoe lang, ze zijn al meteen verrukkelijk.



