Metacognition and Learning (2024) 19:65–101
https://doi.org/10.1007/s11409-023-09349-8

Teachers' judgment accuracy of students' monitoring skills: a conceptual and methodological framework and explorative study

Janneke van de Pol[1] · Sophie Oudman[1]

Received: 28 April 2022 / Accepted: 10 June 2023 / Published online: 18 August 2023

Abstract

Teachers' ability to accurately judge students' monitoring skills is important as it enables teachers to help students becoming better self-regulated learners. Yet, there is hardly any research on this crucial teacher skill and a framework is missing. We present a novel conceptual and methodological framework integrating teachers' judgments of students' monitoring skills with teachers' judgments of students' performance and students' judgments of their own performance. Using this framework, we explored teachers' ability to judge students' monitoring skills and students' performance. Secondary education teachers judged their own students' performance and monitoring on a reading comprehension test ($N_{\text{teachers}} = 46$; $N_{\text{students}} = 406$). Teachers' judgments of students' judgment accuracy deviated 14.33% from the actual accuracy of students' own monitoring judgments, with a tendency to underestimate their students' monitoring accuracy. Teachers' judgments of students' performance deviated 21.96% from students' actual performance, with a tendency to overestimate students' performance. So-called performance cues—pieces of information pertaining to students' prior knowledge and skills relevant to the performance task—appeared predictive or diagnostic both for students' performance and for students' monitoring judgments. When making accurate judgments, teachers used diagnostic cues to a greater extent than when making inaccurate judgments. Yet, when making accurate judgments, teachers also used two non-diagnostic cues (students' IQ and self-concept regarding reading comprehension). To further improve teachers' ability to accurately judge students' monitoring, it may be worthwhile to help teachers ignore non-diagnostic cues.

Keywords Teacher judgments · Teacher monitoring · Monitoring accuracy · Judgment accuracy · Judgments of learning · Metacognitive judgments · Metacognitive skills · Metacognition and teaching · Students' monitoring · Self-regulated learning · Diagnostic competence · Metacomprehension

✉ Janneke van de Pol
j.e.vandepol@uu.nl

[1] Department of Education, Faculty of Social and Behavioural Sciences, Utrecht University, PO Box 80.140, 3508 TC Utrecht, The Netherlands

Accurate monitoring of one's own learning is pivotal for students' academic success (e.g., Dent & Koenka, 2016). When students know well what they do and do not understand, they can regulate their own learning accordingly (Griffin et al., 2019). For example, students can judge that they do not understand a particular text well enough and decide therefore that they need to restudy that text. Accurate monitoring is seen as a prerequisite of effective regulation, which in turn fosters students' learning (Metcalfe & Finn, 2008; Pintrich, 2000; Rawson & Dunlosky, 2007; Thiede et al., 2019; Winne & Hadwin, 1998; Zimmerman, 2000).

However, students have great difficulties with accurate monitoring of their own learning: Students' monitoring judgments of their own performance correlate only weakly with their actual performance (Dunlosky & Lipko, 2007). Interventions using generation tasks such as writing a summary or completing diagrams of relations in a text leads to higher monitoring accuracy (e.g., Prinz et al., 2020; Thiede & Anderson, 2003; Van de Pol et al., 2019, 2020; Van Loon et al., 2014). Yet, students' monitoring is still only moderately accurate when using such interventions.

To further improve students' monitoring skills, teachers could help students with improving their monitoring skills. However, to do this effectively and efficiently, teachers should have an accurate idea of how well students can monitor their own learning. Accurately judging students' monitoring skills can successfully guide teachers' decisions about which students need what kind of help to improve students' monitoring accuracy. Students who underestimate or overestimate their own performance will use inefficient study strategies. Students who underestimate their own understanding may, for example, restudy texts that they already understand whereas students who overestimate their own performance may quit studying a text while not having sufficiently understood previous texts. Instead of taking over students' regulation (e.g., telling students which texts to restudy), helping students to improve their monitoring skills is a more durable solution.

Earlier research has attended to teachers' judgments of students' *performance* (e.g., Thiede et al., 2019; Urhahne & Wijnia, 2021), teachers' judgments of students' *metacognitive abilities* in general (Carr & Kurtz-Costes, 1994), and teachers' judgments of students' *self-regulation strategies* (e.g., Friedrich et al., 2013). Yet, there is hardly any research that focuses on teachers' judgments of students' *monitoring* skills, and more specifically, on how well teachers know how *accurately* students can monitor their own learning. To be able to help students efficiently and effectively in improving their monitoring skills, it is crucial that teachers can signal those students that have trouble accurately monitoring their own learning.

Gaining insight into how accurately teachers can judge students' monitoring skills and how these judgments are established is an important first step in this novel research area. Therefore, we present a framework to conceptualize and empirically study teachers' judgments of students' monitoring skills which can ignite further research on this topic. Furthermore, the results of the current study can give a first impression of teachers' ability to judge students' monitoring skills and can serve as input for ideas on how to improve teachers' judgment accuracy of students' monitoring skills.

Teacher judgments of students' performance

There is a vast body of literature that focuses on teachers' judgments of students' *learning* or *performance* (Südkamp et al., 2012; Urhahne & Wijnia, 2021), rather than teachers' judgments of students' *monitoring skills*. The literature on teachers' judgments of students' performance focuses on three concepts: teacher judgment accuracy, teacher cue-utilization, and cue diagnosticity (Fig. 1, left part). As our framework of teachers' judgments of students' monitoring skills builds on this literature, we will first explain these three concepts in the context of teachers' judgments of students' performance. After that, we will discuss previous research on teachers' judgments of students' monitoring skills, introduce our framework, and discuss teacher cue-utilization and cue diagnosticity in the context of teacher judgments of students' monitoring skills.

Teachers' judgment accuracy of students' performance

Teachers' judgment accuracy of students' performance refers to the relation between a teacher's judgment of a student's performance and a student's actual performance (e.g., Thiede et al., 2019; Van de Pol et al., 2019; Fig. 1; Table 1). There are several measures to express judgment accuracy (Griffin et al., 2019; Thiede et al., 2019). Two common measures are absolute accuracy and bias. These measures focus on the difference between a teacher's judgment of a student's understanding and a student's actual understanding (most often indicated by a test score). If a teacher for example thinks that a student scored eight points on a test and the student indeed scored eight points, the teacher's judgment is perfectly accurate. A teacher judgment of ten points would indicate a deviation of two points (i.e., absolute accuracy) and overestimation of two points (i.e., bias). A teacher judgment of six points would indicate a deviation of two points (absolute accuracy) and an underestimation of two points (i.e., bias). Accurately knowing how well a student understands a particular task (i.e., absolute accuracy/bias) can help teachers to make appropriate and effective instructional decisions (Thiede et al., 2019) and is therefore the focus of this paper (see "Discussion" section for a discussion on whether and how to use this framework for measures of relative accuracy).

Generally, teachers are relatively accurate when monitoring students' learning (Südkamp et al., 2012; Urhahne & Wijnia, 2021). Yet, there is also room for improvement and there are large differences in judgment accuracy between teachers (Südkamp et al., 2012). Although findings differ across studies, teachers appear to have the tendency to overestimate students' performance (Urhahne & Wijnia, 2021). The accuracy of teachers' judgments is assumed to depend on what information or cues they use when making their judgments (e.g., Brunswik, 1952).

Teachers' cue-utilization for judgments of students' performance

When making judgments of students' performance, teachers use all kinds of information or *cues* (e.g., Oudman et al., 2018; Van de Pol 2021a, b) Cues are defined here as "bits of information that might potentially be drawn upon or referred to by a teacher to inform a judgment" (Snow, as cited in Cooksey et al., 2007, p. 431). Teachers use, for example, *performance* cues; cues about students' prior performance on related tasks such as students' performance on a practice task (e.g., Dompnier et al., 2006; Dusek & Joseph, 1983; Helwig et al., 2001; Oudman et al., 2018; Van de Pol et al., 2021a, b). Furthermore, teachers use *student* cues; information that pertains to student characteristics such as their personality, gender, effort in class etc. (e.g., Cooksey et al., 2007; Furnari et al., 2017; Glock et al., 2012). Finally, teachers use *task* cues; information that pertains to the task about which the judgment is made, such as text length, text difficulty, or item difficulty (Cooksey et al., 2007; Oudman et al., 2018).

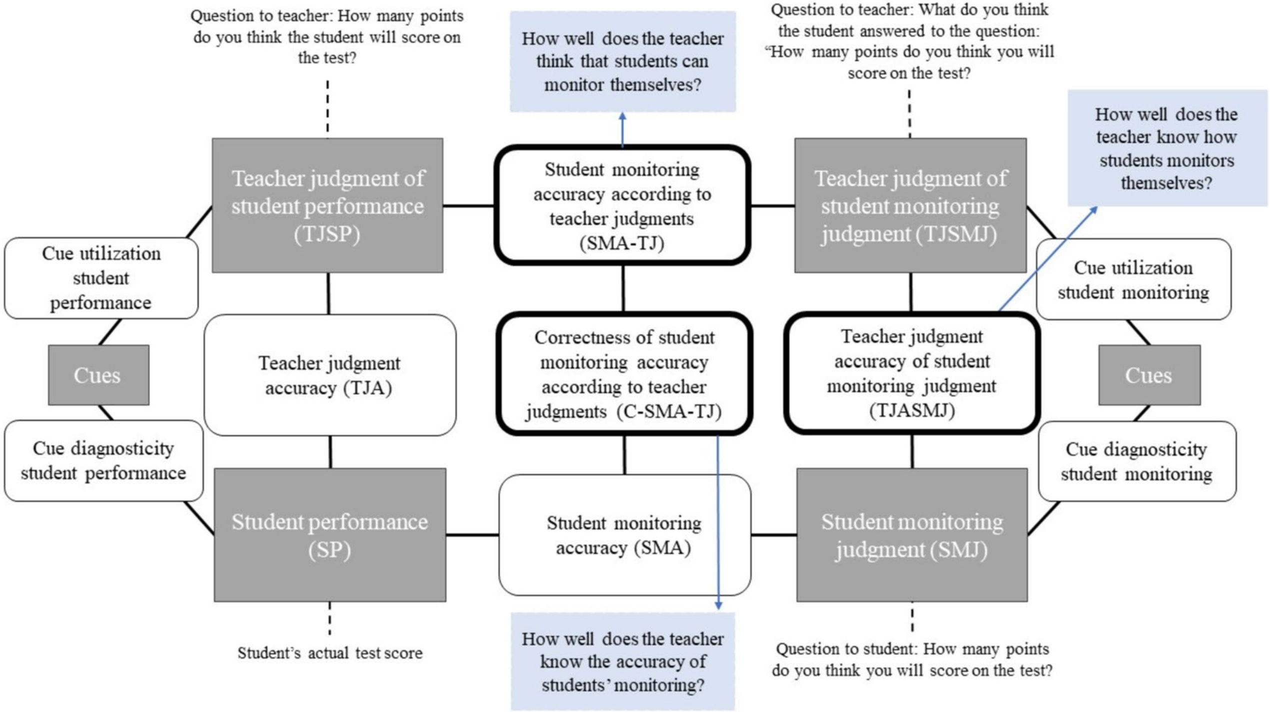
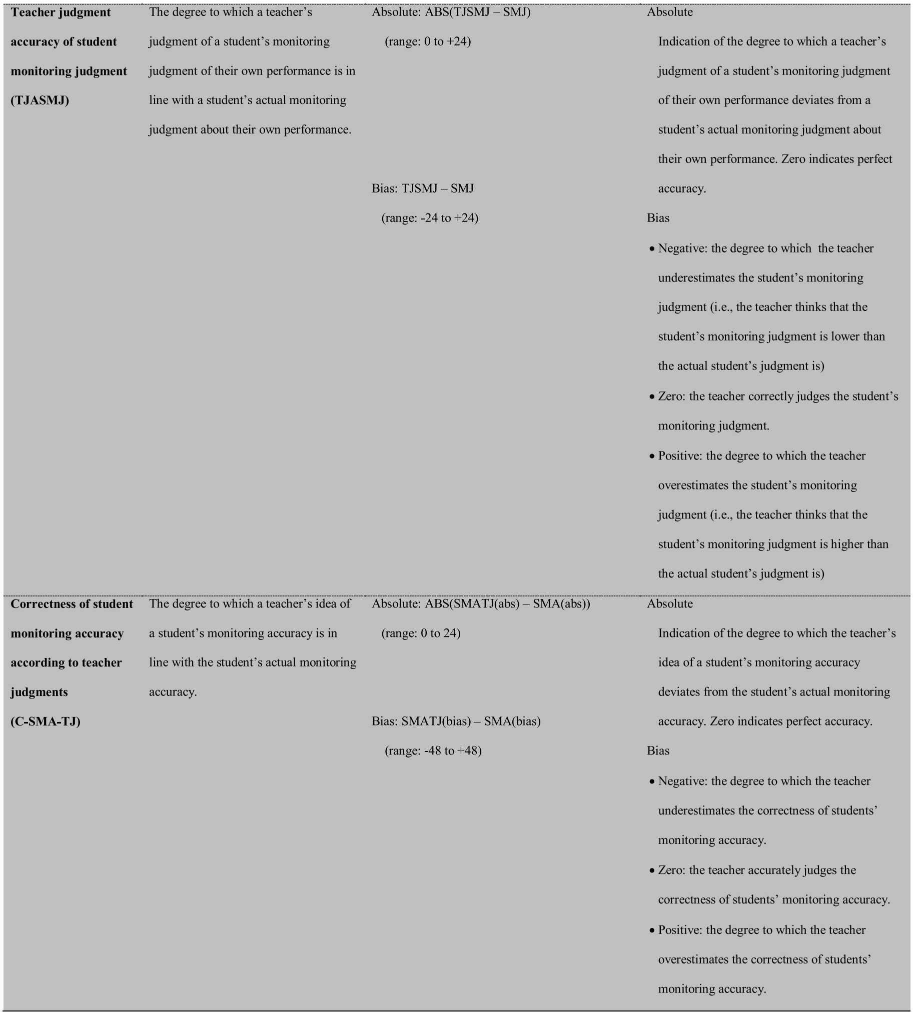

Question to teacher: How many points do you think the student will score on the test?
How well does the teacher think that students can monitor themselves?
Question to teacher: What do you think the student answered to the question: "How many points do you think you will score on the test?
How well does the teacher know how students monitors themselves?
Teacher judgment of student performance (TJSP)
Student monitoring accuracy according to teacher judgments (SMA-TJ)
Teacher judgment of student monitoring judgment (TJSMJ)
Cue utilization student performance
Cue utilization student monitoring
Cues
Teacher judgment accuracy (TJA)
Correctness of student monitoring accuracy according to teacher judgments (C-SMA-TJ)
Teacher judgment accuracy of student monitoring judgment (TJASMJ)
Cues
Cue diagnosticity student performance
Cue diagnosticity student monitoring
Student performance (SP)
Student monitoring accuracy (SMA)
Student monitoring judgment (SMJ)
Student's actual test score
How well does the teacher know the accuracy of students' monitoring?
Question to student: How many points do you think you will score on the test?

Fig. 1 Framework of teachers' judgments of students' monitoring skills. *Note*. Grey boxes indicate constructs that are actually measured. White boxes represent indices that can be derived from the measured constructs. Bold-lined boxes are the variables that are most central to the current study. Blue boxes contain concrete explanations of these central measures that are most central to the current study. Connections between the grey boxes denote the deviation between the constructs represented in the connected boxes

Table 1 Description and explanation of the measures of the framework of teachers' judgments of students' monitoring skills

Measure	Description	Operationalization/calculation	Explanation scores
Student monitoring judgment (SMJ)	A student's judgment of their own performance	Question to student: How many points do you think you will score/have scored on the test?	
Teacher judgment of students' performance (TJSP)	A teacher's judgment of a student's performance	Question to teacher: How many points do you think the student will score/have scored on the test?	
Teacher judgment of student monitoring judgment (TJSMJ)	A teacher's judgment of a student's monitoring judgment	Question to teacher: What do you think the student answered to the question: "How many points do you think you will score/have scored on the test?"	
Student monitoring accuracy (SMA)	The degree to which a student's monitoring judgment about their own performance is in line with their actual performance.	Absolute: ABS(SMJ – SP) (range: 0 to +24) Bias: SMJ – SP (range: -24 to +24)	Absolute The accuracy score indicates the degree of deviation between the student's judgment of performance and actual performance. Zero indicates perfect accuracy. Bias • Negative: the degree to which the student underestimates their own performance • Zero: the student judges their performance accurately • Positive: the degree to which the student overestimates their own performance
Teacher judgment accuracy (TJA)	The degree to which a teacher's monitoring judgment about a student's performance is in line with the student's actual performance.	Absolute: ABS(TJSP – SP) (range: 0 to +24) Bias: TJSP – SP (range: -24 to +24)	Absolute The accuracy score indicates the degree of deviation between the teacher's judgment of the student's performance and the student's actual performance. Zero indicates perfect accuracy. Bias • Negative: the degree to which the teacher underestimates the student's performance • Zero: the teacher judges their performance accurately • Positive: the degree to which the teacher overestimates the student's performance
Student monitoring accuracy according to teacher judgments (SMA-TJ)	The degree to which a teacher's judgment of a student's performance is in line with the teacher's judgment of the student's monitoring judgment of their own performance.	Absolute: ABS(TJSMJ – TJSP) (range: 0 to +24) Bias: TJSMJ – TJSP (range: -24 to +24)	Absolute Indication of the degree to which a teacher's judgment of a student's performance deviates from the teacher's judgment of the student's monitoring judgment. Zero indicates perfect accuracy. Bias • Negative: the degree to which the teacher thinks that the student underestimates their own performance • Zero: teacher thinks that the student can accurately judge their own performance • Positive: the degree to which the teacher thinks that the student overestimates their own performance

Table 1 (continued)



Teacher judgment accuracy of student monitoring judgment (TJASMJ)	The degree to which a teacher's judgment of a student's monitoring judgment of their own performance is in line with a student's actual monitoring judgment about their own performance.	Absolute: ABS(TJSMJ – SMJ) (range: 0 to +24) Bias: TJSMJ – SMJ (range: -24 to +24)	Absolute Indication of the degree to which a teacher's judgment of a student's monitoring judgment of their own performance deviates from a student's actual monitoring judgment about their own performance. Zero indicates perfect accuracy. Bias • Negative: the degree to which the teacher underestimates the student's monitoring judgment (i.e., the teacher thinks that the student's monitoring judgment is lower than the actual student's judgment is) • Zero: the teacher correctly judges the student's monitoring judgment. • Positive: the degree to which the teacher overestimates the student's monitoring judgment (i.e., the teacher thinks that the student's monitoring judgment is higher than the actual student's judgment is)
Correctness of student monitoring accuracy according to teacher judgments (C-SMA-TJ)	The degree to which a teacher's idea of a student's monitoring accuracy is in line with the student's actual monitoring accuracy.	Absolute: ABS(SMATJ(abs) – SMA(abs)) (range: 0 to 24) Bias: SMATJ(bias) – SMA(bias) (range: -48 to +48)	Absolute Indication of the degree to which the teacher's idea of a student's monitoring accuracy deviates from the student's actual monitoring accuracy. Zero indicates perfect accuracy. Bias • Negative: the degree to which the teacher underestimates the correctness of students' monitoring accuracy. • Zero: the teacher accurately judges the correctness of students' monitoring accuracy. • Positive: the degree to which the teacher overestimates the correctness of students' monitoring accuracy.

A series of previous studies, using the same tasks as in the current study, focused on secondary education teachers' cue-utilization when judging students' performance (Van de Pol et al., 2019, 2021a, b). These studies showed that teachers mostly use performance cues but also frequently student cues. In these studies, teachers saw diagrams that their own students completed about the causal relations in the texts the students had read (the study materials for the students were derived from a study of Van Loon et al., 2014). From these diagrams, teachers could deduce several performance cues such as the number of correct relations represented in the diagrams or the extensiveness of the answers in the diagrams. From knowing the student's name, teachers could deduce student cues such as a student's effort in class or conscientiousness when working on school work. Finally, teachers had information about the materials (e.g., texts, test questions) from which they could deduce task cues such as text length or difficulty of a test question. The studies showed that for judgments of students' performance, teachers mostly used performance cues such as the completeness or correctness of the diagrams as well as – although to a lesser extent – student cues (e.g., students' IQ and effort in class) (Van de Pol et al., 2019, 2021a, b). A study of Oudman et al. (2018) showed similar results. In this study—using different materials compared to the current study and the studies of Van de Pol et al., 2019, 2021a, b—primary education teachers judged their students' mathematics performance while having access to students' practice tasks (containing performance cues), students' names (giving access to student cues), and the tasks (giving access to task cues). Teachers mainly used performance cues when these were available: students' answers on the practice task (item performance) and students' misconceptions as shown in the practice tasks.

There are different ways of measuring cue-utilization (Van de Pol et al., 2020). Often, cue-utilization is operationalized by looking at the relation between the actual value or manifestation of a cue (e.g., a student's actual effort level in class or a student's actual score on a prior task) and a teacher's judgment of the student's performance (Fig. 1, far left; Dunlosky et al., 2016; Schleinschok et al., 2017; Van de Pol et al., 2019; Van Loon et al., 2014). For example, teachers who give systematically higher judgments of students' performance on a test for students who score high on a prior task are considered to have used the cue 'prior task score' when making judgments. That is, the significant correlation between teacher judgments of students' performance on a test and the actual cue-value of the cue 'prior task score' indicates that teacher may have used this cue when judging students' performance.

Cue-diagnosticity for students' performance

The cues that teachers use vary in the degree to which they are predictive or *diagnostic* of the outcome, often students' understanding of subject-matter, indicated by students' test performance (e.g., Thiede et al., 2019). A student's score on a prior, related, task will be more diagnostic for a student's test performance than, for example, a student's gender. So cue diagnosticity refers to the relation between an actual cue value and the judged outcome (often: students' test performance; Dunlosky et al., 2016). According to the Lens Model (Brunswik, 1952; Kaufmann & Athanasou, 2009), judgments are more accurate when diagnostic cues are used to arrive at a judgment and when non-diagnostic cues are ignored (Thiede et al., 2015; Van de Pol et al., 2021b). In a previous study (Van de Pol et al., 2021b), the diagnosticity of a wide array of cues for students' reading comprehension performance was investigated; performance cues were found to be most diagnostic (i.e., r> =0.40 and significant; Evans, 1996).

Because we focus on teachers' judgments of students' monitoring skills, not of students' performance, we now turn to this topic. First, we discuss previous research on teachers judgments of students' monitoring skills. Then, we introduce our framework and discuss the concepts of teacher cue-utilization and cue diagnosticity in the context of teacher judgments of students' monitoring skills.

Teachers' judgments of students' monitoring skills

Previous research

A few studies have investigated how teachers judge students' self-regulated learning skills or metacognition. Carr and Kurtz (1991) and Carr and Kurtz-Costes (1994) for example asked primary school teachers to judge students' metacognitive abilities about "where, when, and how strategies can be applied in the classroom" (Carr & Kurtz, 1991, p.267) and these judgments were compared to students' self-reports. Teachers were moderately accurate, as shown by a correlation of .40 between teachers' judgments and students' self-reports (Carr & Kurtz, 1991) and .46 (Carr & Kurtz-Costes, 1994). Using a similar methodology, Friedrich et al. (2013) found weak correlations between secondary school teachers' perceptions of their students' pre-actional self-regulation strategies (e.g., goal setting and planning) and actional self-regulation strategies (e.g., concentration and effort) during math class and students' own perceptions of these aspects (respectively .13 and .29).

These studies, however, used students' self-reports as a criterion, which have been shown to have low validity (Veenman & van Cleef, 2019). In addition, the actual *accuracy* of students' monitoring skills and teachers' ability to judge students' monitoring skills is hardly addressed in previous research. Although it is important for teachers to know whether students know about and use self-regulation strategies, knowing whether students' monitoring is accurate may be crucial as this will help to adapt the regulation strategies teachers use to a student's current self-regulation and will therewith promote students' learning (e.g., Cleary & Zimmerman, 2004; Kostons et al., 2012).

We only encountered two studies that investigated teachers' ability to judge students' monitoring. In a study of Fleury-Roy and Bouffard (2006), primary school teachers ($N=74$) were asked to classify their students ($N=684$) into one of three categories: students who underestimate their own general cognitive abilities ('pessimists'), students who overestimate their abilities ('optimists'), or students who accurately estimate their abilities ('realists'). The teachers' classification was compared to whether students overestimated (classified as optimists), underestimated (classified as pessimists), or accurately (classified as realists) judged their own performance. It turned out that 51.32% of the students were classified correctly by their teachers, meaning that almost 50% of the students were wrongly classified. Of the 112 students who were pessimists, 35 were correctly classified as pessimists by the teachers (31.25%). Of the 88 students who were optimists, 16 were correctly classified by teachers as optimists (18.18%). And the 484 students who were realists, 300 were correctly classified by the teachers as realists (61.98%).

Jamain (2019) conducted a follow-up study asking 13 primary school teachers to classify their students ($N=292$) into the same categories as Fleury-Roy and Bouffard (2006), but separately for students' Mathematics and French performance. Overall, 45.55% (Mathematics) and 41.78% (French) of the students were classified correctly by the teachers. Of the 44 (French) and 47 (mathematics) students who were pessimists, 17 (French) and 20 (mathematics) were correctly classified as pessimists by the teachers (38.64% and 42.55% respectively). Of the 40 (French) and 41 (mathematics) students who were optimists, 10 (French) and 15 (mathematics) were correctly classified by teachers as optimists (25% and 36.59% respectively). Finally, of the 208 (French) and 204 (mathematics) students who were realists, 95 (French) and 98 (mathematics) were correctly classified by the teachers as realists (45.67% and 48.04% respectively).

The categoric approach used in these two studies, however, does not match the actual classroom situation. In this categoric approach, student's monitoring judgment was seen as pessimistic when the students' accuracy score was more than –1 standard deviation from the mean of students' monitoring accuracy, as optimistic when the students' accuracy score was more than + 1 standard deviation from the mean, and as realistic when the students' monitoring accuracy score was within –1/ + 1 standard deviation from the mean.[1] Therefore, the pessimist and optimist group will always be about 16%

[1] Students' monitoring accuracy was calculated by subtracting the students' test score from their monitoring judgment of their test score (both converted to *z* scores).

each, and the realist group about 68%. However, in practice, the amount of students in a class judging their performances optimistic, pessimistic, or realistic can vary from these percentages. The distribution depends on the age of the students, task difficulty, and prior knowledge (Lin et al., 2002). Younger children, for example, overestimate their performance more than older children (Destan & Roebers, 2015). Similarly, students who have low levels of prior knowledge, also tend to overestimate their performance (Dunning et al., 2003; Hacker et al., 2008). In addition, Lin et al. (2002) showed that students' monitoring accuracy depended on the task difficulty; their monitoring was more accurate for easy tasks than for difficult tasks.

Furthermore, the categoric approach of Fleury-Roy and Bouffard (2006) does not enable us to determine the *degree* to which teacher's judgments of students' monitoring accuracy differs from students' actual monitoring accuracy. Yet, knowing this may be important: If teachers misjudge students' monitoring skills with only 1 out of 10 points, there seems to be less of a problem than when teachers misjudge students' monitoring skills with 5 out of 10 points. In the categoric approach, however, both instances (1 or 5 points misjudgment) could be categorized as accurate judgments. Moreover, the categoric approach of Fleury-Roy and Bouffard (2006) does not indicate to what *degree students'* monitoring judgments are correct. This is important as students whose monitoring judgment deviates only slightly from their actual test score may need less or different help than students whose monitoring judgment deviates to a great extent from their actual test score. Yet, in the categoric approach of Fleury-Roy and Bouffard (2006), both these students can be categorized as optimists (or even as realists or pessimists, this is completely dependent of the other students in the sample).

Framework of teachers' judgments of students' monitoring skills

We developed a framework which applies to every classroom situation, regardless of the proportion of optimists/pessimists/realists and which enables us to determine the *degree* to which teachers' judgments of students' monitoring accuracy differs from students' actual monitoring accuracy. This framework can both be used to clearly conceptualize different concepts related to teachers' judgments of teachers' monitoring skills, and as a guideline on how to operationalize these concepts. Teachers' judgment accuracy of students' monitoring judgments is captured with three measures (Fig. 1, bold boxes and Table 1 gray rows).

The first measure is the Student Monitoring Accuracy according to Teacher Judgments which indicates how accurately a teacher thinks students can monitor their own understanding (SMA-TJ; Fig. 1 and Table 1). This measure refers to the (absolute or signed) difference between a Teacher's Judgment of a Student's Performance (TJSP) and a Teacher's Judgment of a Student's Monitoring Judgment (TJSMJ). Teachers' Judgments of Students' Performance (TJSP) can for example be measured by asking teachers: 'How many points do you think the student will score (or has scored) on the test?' (Fig. 1). Teachers' Judgments of Students' Monitoring Judgments (TJSMJ) can be measured by asking teachers 'What do you think the student answered to the question: How many points do you think you will score on the test?' (Fig. 1). This measure thus shows to what degree a teacher's judgment of a student's performance is in line with the teacher's judgment of the student's monitoring judgment, regardless of the degree to which the teacher is correct about the student's actual monitoring judgment.

The difference between teachers' Judgments of Students' Monitoring Judgments (TJSMJ) and Student Monitoring Judgments (SMJ) is represented by the second measure: the Teacher Judgment Accuracy of Students' Monitoring Judgment (TJASMJ; Fig. 1 and Table 1). This measure indicates the degree to which a teacher's judgment of a student's monitoring judgment is in line with a student's actual monitoring judgment. Or in other words: How well does the teacher know how students monitors themselves? A difference of zero indicates that the teacher can accurately judge a student's monitoring judgment; a negative difference indicates underestimation of the student's monitoring judgment (i.e., the teacher thinks that the student underestimates their performance more than the student actually does) and a positive difference an overestimation (i.e., the teacher thinks that the student overestimates their performance more than the student actually does). These two first measures, the Student Monitoring Accuracy according to Teacher Judgments (SMA-TJ) and the Teacher Judgment Accuracy of the Student's Monitoring Judgment (TJASMJ), do not indicate how well teachers are able to judge their students' monitoring judgment accuracy. This is indicated by the third measure.

The third measure, the Correctness of Student Monitoring Accuracy according to Teacher Judgments (C-SMA-TJ; Fig. 1/Table 1), compares the difference between Student Monitoring Accuracy according to Teachers' Judgments (SMA-TJ) and a Student's actual Monitoring Accuracy (SMA; i.e., the difference between a student's monitoring judgment [SMJ] and their Student's actual Performance [SP]). It indicates how well a teacher knows how accurately a student can monitor their understanding. For this measure (C-SMA-TJ), a difference of zero means that that the teacher made an accurate judgment of the students' degree of monitoring accuracy; a negative difference indicates underestimation of the student's monitoring accuracy and a positive difference an overestimation.

We will further clarify the framework using a numeric example (Fig. 2). In this example, the student scored two points on a test (Student Performance or SP = 2). The student judged that they would score four points (Student Monitoring Judgment or SMJ = 4). The student's monitoring judgment thus deviates two points from their actual performance (Student Monitoring Accuracy or $\text{SMA}_{\text{absolute accuracy}} = 2$) and the student overestimates their performance with two points (Student Monitoring Accuracy or $\text{SMA}_{\text{bias}} = +2$).

The teacher thinks that the student scored five points (TJSP). Given that the student's actual performance (SP) was two, the teacher's judgment of the student's performance deviates three points from the student's actual performance ($\text{TJA}_{\text{absolute accuracy}} = 3$) and the teacher overestimates the student's performance with three points ($\text{TJA}_{\text{bias score}} = +3$).

Furthermore, the teacher thinks that the student's monitoring judgment was two points (TJSMJ = 2). Given that the teacher thought that the student would score five points, the s*tudent monitoring accuracy according to teacher judgments* is three points ($\text{SMA-TJ}_{\text{absolute accuracy}} = 3$) and the teacher underestimates the student's monitoring judgment with three points ($\text{SMA-TJ}_{\text{bias score}} = –3$).

In reality, the student thought they would score four points (SMJ), so the *teacher judgment of the student's monitoring judgment* (TJSMJ) deviates two points from the student's monitoring judgment ($\text{TJASMJ}_{\text{absolute accuracy}} = 2$) and the teacher underestimates the student's monitoring judgment with two points ($\text{TJASMJ}_{\text{bias score}} = \text{-}2$).

Finally, the teacher thinks that the student underestimates their own performance by three points (SMA-TJ) while the student overestimates their performance by two points (SMA). Therefore, the teacher's idea of the student's monitoring accuracy deviates one point from the student's actual monitoring accuracy ($\text{C-SMA-TJ}_{\text{absolute accuracy}} = 1$) and the teacher

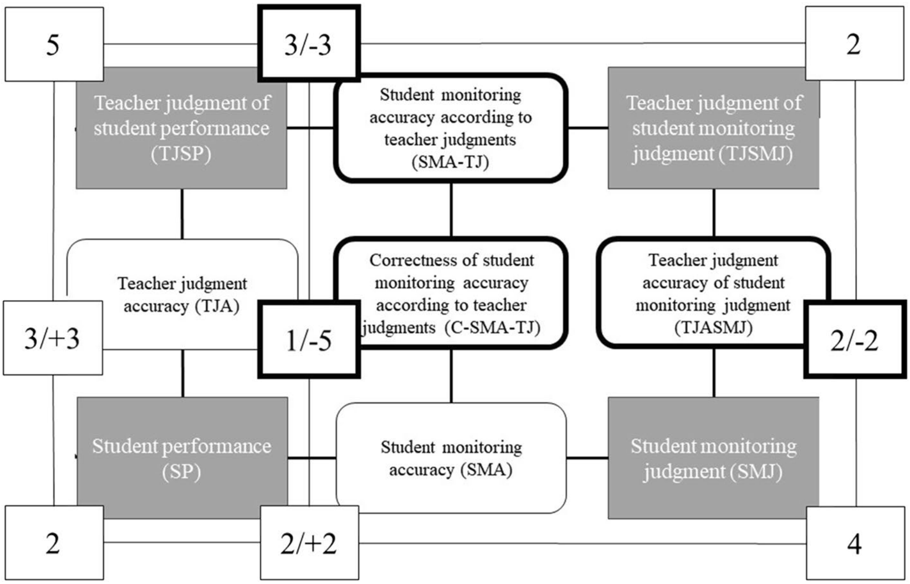

5
3/-3
2
Teacher judgment of student performance (TJSP)
Student monitoring accuracy according to teacher judgments (SMA-TJ)
Teacher judgment of student monitoring judgment (TJSMJ)
Teacher judgment accuracy (TJA)
Correctness of student monitoring accuracy according to teacher judgments (C-SMA-TJ)
Teacher judgment accuracy of student monitoring judgment (TJASMJ)
3/+3
1/-5
2/-2
Student performance (SP)
Student monitoring accuracy (SMA)
Student monitoring judgment (SMJ)
2
2/+2
4

Fig. 2 Numeric example for the framework of teachers' judgments of students' monitoring skills. *Note.* First number represents the absolute accuracy score, second number the bias score

underestimates the correctness of the student's monitoring accuracy (C-SMA-TJ$_{\text{bias}}$ = -5). That is, the teacher unjustly thinks that the student underestimates their understanding.

Thus, the *student monitoring accuracy according to teacher judgments* (SMA-TJ) indicates the degree to which the teacher *thinks* the student accurately monitor their own performance. With the *teacher judgment accuracy of the student's monitoring judgment* (TJASMJ), one can check whether the teacher's idea of a student's monitoring judgment is correct. Finally, the *correctness of student monitoring accuracy according to teacher judgments* (C-SMA-TJ) indicates the extent to which the teacher's idea of a student's monitoring accuracy is in line with the student's actual monitoring accuracy.

Teachers' cue-utilization for judgments of students' monitoring judgments

As with judging students' performance, teachers also use cues when judging students' monitoring judgments. These cues also vary in the extent to which they are diagnostic of the outcome that is being judged; in this case, students' monitoring judgments (Fig. 1, far right). So for judging students' monitoring judgments, cue-utilization refers to the relation between the manifestation or value of a cue (e.g., a student's intelligence) and a teacher's judgment of a student's monitoring judgment.[2] If teachers give systematically higher judgments of students' monitoring judgments for students who are highly conscientious (as perceived by the teachers), there is a (positive) relationship between the (perceived) cue value and the teachers' judgments, indicating that teachers may have used this cue.

[2] Some studies use the actual cue values for this cue-utilization measure (e.g., Van Loon et al., 2014); other studies use the cue values as perceived by the teacher (e.g., Fleury-Roy & Bouffard, 2006).

Little is known about the cues teachers use for judging students' monitoring skills. Dignath and Sprenger (2020) asked primary school teachers how they would asses students' self-regulated learning. From their answers, it occurred that teachers used students' off-task behavior and self-assessment as indicators for poor self-regulation. When being asked to identify students with poor self-regulation, teachers in the study of Callan and Shim (2019) indicated to focus on off-task behavior and poor academic performance. Finally, Carr and Kurtz-Costes (1994) found that teachers' used students' self-concept and achievement level as primary indicators of students' metacognitive ability. Summarizing, when judging students' self-regulation, teachers seem to use students' off-task behavior, self-concept, and (self-assessed) achievement. In the current study, we investigate teacher's cue-utilization when judging students' monitoring skills and empirically determine the actual diagnosticity of the cues.

Cue-diagnosticity for judgments of students' monitoring

Teachers' judgments of students' monitoring judgments are based on cues. The diagnosticity of these cues refers to the relation between a student's monitoring judgment of their test score and the value of a cue. If, for example, students' monitoring judgments strongly relate to the degree of students' conscientiousness (i.e., a cue), this cue is considered diagnostic for students' monitoring judgments. In line with the assumptions of the Lens Model (Brunswik, 1952), teachers' use of cues that are diagnostic for students monitoring judgments would lead to more accurate teacher judgments of students' monitoring judgments. That is, if teachers use the same cues (for their judgment of a student's monitoring judgment) as students (for judging their performance), they may arrive at similar judgments as students. In this sense, the diagnosticity of cues for students' monitoring judgments is similar to students' cue-utilization for judging their own performance. When judging their performance, students tend to rely on non-diagnostic cues such as text length, interest in the text topic, or study effort instead of diagnostic performance cues such as their ability to explain the meaning of a text, resulting in inaccurate judgments (Bol et al., 2010; Hacker et al., 2008; Thiede et al., 2010). Thus, when teachers use the same non-diagnostic cues, they probably arrive at accurate judgments of students' monitoring judgments. Hence, the cues that teachers use should differ between their judgments of students' monitoring and judgments of students' performance to arrive at accurate judgments.

The current study

In the current study, we focused on secondary education teachers' judgments of their own students' monitoring. Teachers judged both students' performance (TJSP) on a reading comprehension test and students' monitoring judgments (SMJ) about their own reading comprehension performance (TJSMJ). While making the performance judgments, all teachers had the following information from which they could deduce cues: (1) student work (i.e., diagrams students completed about the causal relations in the text) making performance cues available, (2) students' names, making student cues available, and (3) the task (i.e., the texts students have read and the test questions about these texts), making task cues available. Students also gave a monitoring judgment of their own test performance (SMJ) and took a test to measure student performance (SP).

Our first research question was:

1. To what extent:

   a) do teachers think students can accurately monitor their own learning? (student monitoring accuracy according to teacher judgments, SMA-TJ)
   b) are teachers' judgments of students' monitoring judgements in line with students' actual monitoring judgments (teacher judgment accuracy of student monitoring judgment, TJASMJ)?
   c) are teachers' judgments of students' monitoring accuracy in line with students' actual monitoring accuracy (correctness of student monitoring accuracy according to teacher judgments, C-SMA-TJ)?

Given a lack of previous research, we did not have hypotheses for RQ1a – RQ1c; these aspects were explored. For aforementioned reasons (not matching the actual classroom situation and not indicating the degree of judgment (in)accuracy, we do not use the categoric approach of Fleury-Roy and Bouffard (2006) in this paper. Yet, to be able to compare teachers' judgment accuracy of students' monitoring skills in our current sample to that of the samples used in Fleury-Roy and Bouffard (2006) and Jamain (2019), we do provide the categoric calculations as applied to our data in the Supplementary Material.

To arrive at accurate judgments of students' performance and students' monitoring judgments, teachers need to use those cues that are most diagnostic for students' performance and students' monitoring judgments respectively. Previous research showed that performance cues were most diagnostic for students' reading comprehension performance (Van de Pol et al., 2021b). Yet, for teachers' judgments of students' monitoring judgments, it is unclear which cues are diagnostic. In the current study, we will therefore explore the diagnosticity of several cues for judging students' monitoring judgments and compare this to the diagnosticity of cues for students' performance. Our second research question is:

2. What cues are diagnostic for students' monitoring judgments and how does this differ compared to the diagnosticity of cues for students' performance?

For judgments of students' monitoring judgments we expect those cues that students' use (such as text length or interest in the text topic) to arrive at their monitoring judgments of their performance to be most diagnostic for students' monitoring judgments. Based on previous research we expect that students use mostly non-diagnostic (student or task) cues (Bol et al., 2010; Hacker et al., 2008; Thiede et al., 2010), and these cues are thus expected to be most diagnostic for students' monitoring judgments.

Finally, we explored teachers' cue-utilization both for judging students' performance and students' monitoring judgments. Furthermore, we compared cue-utilization for accurate teacher judgments of students' performance or students' monitoring judgments to inaccurate teacher judgments. For cue-utilization, we addressed the following research questions:

3. What cues do teachers use for judging students' performance (RQ3a) and to what extent does this differ between accurate and inaccurate judgments of students' performance (RQ3b)?
4. What cues do teachers use for judging students' monitoring judgments (RQ4a) and to what extent does this differ between accurate and inaccurate judgments of students' monitoring judgments (RQ4b)?

We expect that for teachers' judgments of students' performance (RQ3a), teachers will mostly use performance cues (e.g., the completeness and correctness of students' diagrams) as well as—although to a lesser extent—some student cues (e.g., IQ and effort in class) (Van de Pol et al., 2019, 2021a, b). For accurate judgments, we expect teachers to use performance cues to a greater extent than for inaccurate judgments (RQ3b).

Regarding teachers' judgments of students' monitoring judgments (RQ4a), we expect –based on research on teachers' judgments on students' self-regulation (Callan & Shim, 2019; Carr & Kurtz, 1994; Dignath & Sprenger, 2020)—that teachers use cues regarding students' academic performance (in our study: cues from the diagrams students' have completed, students' general school performance), students' behavior in class (here e.g., students' effort in class and conscientiousness), and students' self-concept.

In addition (RQ4b), based on the Lens Model (Brunswik, 1952), we expect that for accurate judgments, teachers use the diagnostic cues to a greater extent than for inaccurate judgments (the diagnosticity is determined under RQ2).

Method

Participants and design

Forty-six secondary education teachers of different subjects for which text comprehension is important (e.g., languages, history) participated (65.22% female; 95.65% Dutch; $M_{\text{age}} = 40.90$; $SD = 0.84$). They made judgments about nine of their own students, resulting in a total sample of 406 students ($M_{\text{age}} = 15.15$ years, $SD = 1.34$; 52.71% female; 90.14% Dutch).[3] The sample-size was based on a multilevel a-priori power analysis (power = 0.80) conducted in spa-ml (Moerbeek & Teerenstra, 2015). Teachers were recruited from the researcher's network and via social media and received a €50 voucher for participation. Teachers were asked to indicate with which classes they could participate potentially and one of those classes was selected randomly. A 10€ voucher was put up for raffle within each participating class. The study was approved by the Ethics committee of the authors' institute.

The data for the current study stems from a larger study addressing, amongst others, the relationship between teachers' monitoring accuracy of students' performance and their cue-utilization for these performance judgments (Van de Pol et al., 2021b). For the purpose of this larger study, teachers participated in three conditions. First, teachers made

[3] Two teachers made judgments about 7 students and one teacher about 5 students because there were not enough students in their class that participated in this study due to non-consent or illness on the day of administration.

performance judgments, restudy decisions for students, and judgments about students' monitoring judgments for one practice student and three of their students (judgment-only condition). Second, they made performance judgments, indicated on a cue-list which cues they had used for these performance judgments, made restudy decisions, and made judgments about students' monitoring judgments for one practice student and three of their students (judgment + cue-list condition). Third, they made performance judgments, indicated on a cue-list which cues they had used for these performance judgments, judged the values of the cues they had used, made restudy decisions, and made judgments about students' monitoring judgments for one practice student and three of their students (judgment + cue-list + cue-value-judgment condition). This order (judgment-only condition, judgment + cue-list condition, judgment + cue-list + cue-value-judgment condition) was the same for each teacher. In all cases, teachers saw the diagrams that students completed (making performance cues available), the student's name (making student cues available), and they knew the texts and the test questions (making task cues available). The conditions did not differ regarding the three accuracy measures[4] (SMA-TJ, C-SMA-TJ, and TJASMJ). Therefore, in the current study, we used data of all three conditions.

To ensure variation in students' reading comprehension level, students per condition were selected based on their scores on the *general reading comprehension level* test (see "Measurement of student cues" section). For each condition, we selected a student that scored around the 20th, 50th, and 80th percentile within each class and the order in which the students had to be judged within a condition was randomized. The practice students were chosen randomly from each teacher's class.

Materials

Expository texts

Students read three expository texts "Music makes smart" (167 words; *M* reading time 89.40 s, *SD* = 135.22), "Sinking of metro cars" (158 words; average reading time 124.70 s; *SD* = 64.18), and "Concrete constructions" (166 words; average reading time 72.86 s; *SD* = 47.26), stemming fromVan Loon et al. (2014).[5] Each text contained four causal relations between five elements (e.g., because the concrete dries out (element 1), the building becomes smaller (element 2); 1 causal relation). For each text, students received the following instructions: "Please read this text carefully. You cannot look back in the text when you will complete diagrams and take the test." (Van de Pol et al., 2021b). There was no time limit for reading the texts. After having read each text, students indicated how many words in the text they found difficult, to measure the task cue 'number of difficult words in the text'.[6]

[4] To check this, we performed multilevel analyses with students nested in teachers using a Bonferroni corrected alpha of .01 for multiple comparisons. Condition did not have an effect on SMA-TJ, C-SMA-TJ, and TJASMJ (*p*'s between .023 and .900).

[5] This is also the order in which the students read the texts and teachers judged students' performance (TJSP) and judged students' monitoring judgments (TJSMJ).

[6] Other task cues that were measured were: number of facts in each text, text length, and text position. However, we did not include these cues in the current study because these did not vary per text, so cue-diagnosticity and cue-utilization cannot be computed for these cues.

Diagrams

After having read all texts, students completed three pre-structured diagrams, one for each text (Van de Pol et al., 2021b; Van Loon et al., 2014). There was no time limit for completing the diagrams. Each diagram consisted of five boxes connected with arrows and one box was already completed (see Appendix Fig. 4 for an example). Upon the completion of each diagram, students received the following instructions: "Please complete the diagram for the text [title text] that you have read. If you are unable to complete a box, please fill out a ?". To determine the values of several diagram cues, students' diagrams were coded using an existing coding scheme (Van de Pol et al., 2021b; Van Loon et al., 2014). Several performance or diagram cues were coded. We coded students' text in the diagrams as a correct element or box (see Appendix Fig. 4) when the student's answer matched the answer in the coding scheme, as a commission error when the student's answer was not in the coding scheme, or as an omission when a necessary cause or consequence (i.e., an element) from the coding scheme was missing in the diagram. In addition, we counted the number of question marks or blank boxes in the diagram. In addition, the number of correct relations in the diagram was determined; that is, the number of correct combinations of two elements or boxes (see Appendix Fig. 4). The minimum for these diagram cues was 0 and the maximum is 4, per diagram (0–12 for the three diagrams together).

In addition, the number of correct facts (0–7) that each diagram contained was scored (i.e., text details that were not essential for understanding the causal relations). Two assistants coded 60 diagrams and the interrater reliability was considered substantial for coding the correct elements, commission errors, omissions, number of empty boxes and number of question marks in the diagrams (Krippendorff's alpha = 0.96; Landis & Koch, 1977), for determining the correct combinations of two elements (Krippendorff's alpha = 0.91) and for coding the facts (Krippendorff's alpha = 0.99). The average number of words per diagram box was determined and the time needed to finish each diagram was extracted from the online platform in which the data was gathered (i.e., Gorilla).

Measurement of student cues

Students reported their *gender* and whether they had a *learning problem. Students' general effort in class* was measured using the Ongoing Engagement Subdomain Scale (IRRE, 1998). Students answered five questions on a scale from 1 (totally disagree) to 4 (totally agree). An example item is: "I pay attention in the lessons of teacher X". The internal consistency in our sample was acceptable ($\Omega = 0.76$).

Students'conscientiousness and *extraversion* was measured using the Big Five conscientiousness and extraversion scales (Goldberg, 1992). Students answered six questions for conscientiousness and six for extraversion on a scale from 1 (not true at all) to 7 (entirely true). Example items are: "To what extent do you show the following traits in class of teacher X: precision (for conscientiousness) or quietness (reverse coded; for extraversion)". The internal consistency in our sample was good ($\Omega = 0.86$ for conscientiousness, $\Omega = 0.89$ for extraversion).

Students were asked to report their average grade for the core subjects Dutch, mathematics, geography, science, biology, and English. The mean of these grades was used for the variable *general school performance.*

Students' interest in each text topic was measured using the Situational Interest Scale (Linnenbrink-Garcia et al., 2010), that was administered separately for each text. Students answered four questions on a scale from 1 (not at all true) to 5 (very true). An example item is: “The topic of this text is fascinating to me”. The internal consistency in our sample was good ($\Omega = 1.00$).

Students' self-efficacy for the subject their teacher teaches to them was measured using the Perceived Self-efficacy Scale (Marsh et al., 2006). Students answered four questions on a scale from 1 (almost never) to 4 (almost always). An example item is: “I'm certain I can. understand the most difficult materials presented in the study materials of subject X”. The internal consistency in our sample was good ($\Omega = 0.83$).

Students' self-concept for reading comprehension was measured using the self-concept scale of the TIMSS 2007 study, that was adapted for reading comprehension (Olson et al., 2008). Students answered four questions on a scale from 1 (totally disagree) to 4 (totally agree). An example item is: “I usually do well in reading comprehension tasks”. The internal consistency in our sample was good ($\Omega = 0.88$).

Students' nationality was measured by asking students' own birth country, that of their father, and mother. A variable consisting of the following scores was then computed: 0 (student/father/mother not born in the Netherlands (NL)); 1 (student not born in NL; mother or father born in NL), 2 (student not born in NL; mother and father born in NL), 3 (student born in NL; mother and father not), 4 (student and mother or father born in NL), or 5 (student, mother and father born in NL).

We used the following three quality criteria to assess the quality of the instruments used to measure the student cues with regard to *knowledge and understanding* (i.e., general reading comprehension, reproduction of facts, prior knowledge, students' IQ): 1) reliability (here $\Omega \geq 0.70$), 2) question difficulty (% correct for open questions 25–90%; for MC questions we use the corrected p-value which indicates the percentage of students who can answer the question correctly without guessing[7]), and 3) discrimination (item-rest correlations of ≥ 0.25; Van Berkel & Bax, 2006; Van den Brink & Mellenbergh, 1998). Instruments that scored insufficiently on two or more of these three quality criteria (i.e., reliability, question difficulty, and discrimination) were not used (here: prior knowledge test about the test topic).

Students' *general reading comprehension level* was measured using a cloze test (Kamalski, 2007), that was developed for this project. In this cloze test, students had to read a text and complete 20 words that were omitted in the text. See for further information about this test Van de Pol et al. (2021b). The items were of mixed difficulty but not too difficult or too easy: the percentage correctly answered items ranged from 9.7% to 92% ($M = 62\%$; $SD = 22\%$; Van Berkel & Bax, 2006; Van den Brink & Mellenbergh, 1998). The item-rest correlations were sufficient for 14 of the 20 items ($M = 0.18$; $SD = 0.08$): the majority of the items thus discriminated well between students with low and high test scores (Van Berkel & Bax, 2006; Van den Brink & Mellenbergh, 1998).

Students' IQ was measured using the shortened version of the Raven Progressive Matrices (Bilker et al., 2012), containing nine items. The internal consistency in our sample was moderate ($\Omega = 0.54$). The item difficulty varied and items (except for three) were not too easy nor too hard (corrected p-values for the items that scored sufficiently varied between 0.63 and 0.87; threshold was 0.56). The item-rest correlations were sufficient to very good ($M = 0.27$; $SD = 0.07$). See also Van de Pol et al. (2021b).

[7] Probability p = proportion of students that answered the question correctly. $P_{corrected} = p - (1-p)/(\text{number of alternatives}-1)$.

To measure s*tudents' ability to reproduce facts*, students first read a text from Van Loon et al. (2014). After that, they completed the personality questionnaire for conscientiousness and extraversion. Thereafter, they (without seeing the text they previously read) answered five questions about the facts in the text. Answers of 90 students were double coded and the interrater reliability was good (Krippendorff's alpha = 0.98). The items were not too difficult or too easy: the percentage correctly answered items ranged from 27.8% to 78% with one exception: that question had a percentage correct of 18.8% which was slightly below the threshold of 25%. The item-rest correlations were sufficient for four of the five items (M = 0.17; SD = 0.002). See also Van de Pol et al. (2021b).

Student text comprehension test

After finishing the diagrams, participants answered one question per text that asked them to describe the four causal relations from the text using the five elements (Van de Pol et al., 2019; Loon et al., 2014). One element was always provided in the formulation of the test question (see Appendix for an example test question). Students' tests were coded using an existing coding scheme (Van de Pol et al., 2019; Van Loon et al., 2014). Students could score – per text – one point for each correct element (four in total; the event that was provided in the test question was not counted) and one point for each correct relation between two element (four in total; the element that was provided in the test question was counted here). The overall minimum score was 0 and the maximum score was 24.[8] Data of 50 students was coded by two assistants. The interrater reliability was substantial both for scoring the elements (Krippendorff's alpha = 0.93) and for scoring the causal relations (Krippendorff's alpha = 0.88). The reliability of the test was acceptable (Cronbach's alpha = 0.73).[9]

Student Monitoring Judgment (SMJ)

To measure students' monitoring judgments (SMJ) of their overall test score (so over the three test questions), we asked the following question: "The maximum number of points for the test is 24: 12 for the correct elements (causes or consequences) and 12 for the right order of the elements. How many points do you think you scored on the test?".

Teacher Judgments of Student Performance (TJSP)

As an introduction for the question to measure teachers' judgments of student performance (TJSP) per test question, teachers received the following explanation: "In a bit, you will judge for a few students how many points they have scored on each test question. For each test question, students could score a maximum of eight points: four points for the four correct elements (causes or consequences), and four points for the four correct causal relations (i.e., the correct combination of two elements). You can use the following sources when making these judgments: the diagram the student completed, information you know about your student, and information that you remember from the texts and test questions." Then, teachers were asked, per text: "How many of the eight points do you think this student scored on the test question about [TEXT TITLE]?". As the other judgments were

[8] 3 (texts) * (4 (correct elements) + 4 (correct relations)) = 24 points.

[9] For more information about the test and its validity, see Van de Pol et al. (2021b).

asked at the test level and not per text, we used the sum score of the teachers' judgments of students' performance over the three texts in the analyses.

Teacher Judgment of Students' Monitoring Judgment (TJSMJ)

To measure teachers' judgments of students' monitoring judgments, we asked teachers the following question: "For the entire test (about the three texts), students could score 24 points in total: 12 for the correct elements (causes or consequences), and 12 for the correct causal relations (i.e., the correct combination of two elements). How many points do you think that each student themselves think they have scored on the test? The students have judged their own test scores after they had completed the test."

Procedure

Students

In a first session (± 50 min), students completed several tasks and questions to measure the values of the cues we used in this study (see sections about Diagram cues and Student cues). In addition, students completed other measures that are not included in the current study such as the liking of and familiarity with peers.

In a second session which took place one week later and took ± 50 min, students watched a short instruction movie clip explaining the tasks, read the three texts, and indicated per text how many words they found difficult and how interested they were in the text topic. Then, they completed a pre-structured diagram about the causal relations in the text (without seeing the texts). After finishing the three diagrams, they completed the three test questions (without seeing the texts or diagrams). Finally, they judged their own and the understanding of five of their peers.[10] Students completed all tasks of session one and two on a computer while being in class; the sessions were led by a researcher. All tasks were programmed in Gorilla.

Teachers

First, teachers read three texts. Then, they judged – per test question – how many points they thought a particular student scored on the test (TJSP), made restudy decisions for this student,[11] and judged how many points they thought the student thought they scored on the test (TJSMJ). Teachers started with judgments about one practice student (from their own class) to get accustomed to the procedure. This practice student was selected randomly from the teachers' class. The data on the practice student were not included in analyses. Then, they made the judgments for nine of their own students. The session lasted for about 60 to 90 min. Teachers completed all tasks on a computer with a researcher present in case of questions. All tasks were programmed in Gorilla.

Analyses

See Table 1 for the calculations used to measure the main accuracy measures of the framework (SMA-TJ, TJASMJ, and C-SMA-TJ). When providing descriptive statistics, we provide these

[10] The peer judgments fall outside the scope of this article.

[11] The restudy decisions fall outside the scope of this article.

for both absolute accuracy and the bias measure of accuracy. When conducting analyses, we only use absolute accuracy because negative and positive values of bias can cancel each other out. The correlations between all cues can be found in Table S2 in the Supplementary Material.

Analyses RQ1

Using our framework, we provide descriptive statistics (*M, SD*) regarding the degree to which: (RQ1a) teachers think students can accurately monitor their own learning (SMA-TJ), (RQ1b) teachers' judgments of students' monitoring judgements are in line with students' actual monitoring judgments (TJASMJ), and (RQ1c) teachers' judgments of students' monitoring accuracy are in line with students' actual monitoring accuracy (C-SMA-TJ). For the sake of completeness, we also report teachers' judgment accuracy of students' performance (TJA) and students' monitoring accuracy (SMA).

Analyses RQ2, 3, and 4

RQ2, RQ3, and RQ4 were answered by performing multilevel regression analyses in Mplus version 8 (Muthén & Muthén, 1998–2017) using the maximum likelihood estimation with robust standard errors (MLR) which is robust to non-normality. All predictor variables were centered around the grand mean. We used the "complex" function to account for the nested data structure with students (level 1) clustered within teachers (level 2). The proportion of variance at the student and the teacher level for each accuracy measure can be found in Table S3 in the Supplementary Material.

Analyses for RQ2 about cue diagnosticity For the current study, we were interested in the extent to which a wide range of cues (Table 3) was diagnostic for students' performance[12] and students' monitoring judgments. We determined the diagnosticity of the cues for students' performance by regressing students' performance (i.e., their total test scores) on the cue values (e.g., students' IQ score; Van de Pol et al., 2021b). Likewise, to determine the diagnosticity of the cues for students' monitoring judgments, students' monitoring judgments were regressed on the cue values. The standardized regression coefficients indicated the average diagnosticity score for the whole sample. High correlations (negative or positive) indicate high diagnosticity.

To compute the diagnosticity per cue, we used single regression models (one model for each cue) because we were interested in "total diagnosticity" of each cue, that is, the explained variance in students' performance and monitoring judgments by each cue, including shared explained variance by other cues we measured. Additionally, we explored the amount of variance in students' performance and students' monitoring judgments that was explained by all cues together, using multiple regression analyses including all cues.[13]

Analyses for RQ3 and 4 about cue-utilization To determine teachers' cue-utilization for their judgments of student performance (TJSP), teachers' judgments of student

[12] The diagnosticity of these cues for students' performance is also reported in Van de Pol et al. (2021b). To be able to compare the diagnosticity of cues for students' performance with the diagnosticity of the cues for students' monitoring judgments, we will also report the diagnosticity of the cues for students' performance here.

[13] Multicollinearity was not an issue because we were only interested in the total amount of explained variance of all cues together.

performance were regressed on the cue values (e.g., students' IQ scores; cf. Van Loon et al., 2014). Likewise, to determine teachers' cue-utilization for their judgments of students' monitoring judgments, teachers' judgments of students' monitoring judgments (TJSMJ) were regressed on the cue-values (e.g., students' IQ scores). High standardized regression coefficients (negative or positive) indicate high usage.

The single regression models (one model for each cue) that we used to analyze teachers' cue-utilization, indicate teachers' "maximal usage" of each cue, that is, the explained variance in teachers' judgments by each cue, including shared explained variance by other cues we measured. Additionally, we explored the amount of variance in teachers' judgments of students' performance (RQ3) and teachers' judgments of students' monitoring judgments (RQ4) that was explained by all cues together, using multiple regression analyses including all cues.[13]

In addition, we tested whether cue-utilization differed between: accurate ($n=78$) and inaccurate ($n=80$) teacher judgments of students' *performance* (RQ3b), and between accurate ($n=79$) and inaccurate ($n=76$) teacher judgments of students' monitoring judgments (RQ4b). We defined accurate judgments as teacher judgment accuracy scores (both teacher judgment accuracy of students' performance [TJA] and of students' monitoring judgments [TJASMJ]) as scores that deviate ≤ 1 standard deviation from the mean (+ or -) and inaccurate judgments as scores that deviate more than 1 standard deviation from the mean (+ or -).

More specifically, we tested the effects of the interaction terms between group (accurate/ inaccurate) and each cue (e.g., students' self-concept regarding reading comprehension, students' IQ) on teachers' judgments of students' performance (RQ3b) or students' monitoring judgments (RQ4b). Cues that had a significant correlation of ≥ (-)0.40 were considered diagnostic or (probably) used (Evans, 1996). We only analyzed the interaction effects for RQ3b and RQ4b for those cues that had significant correlations of ≥ (-)0.40 for cue-use in the whole sample, highly accurate sample, or highly inaccurate sample.

Outliers

For each variable that was used in the analyses of RQ2, RQ3, and RQ4 we identified univariate outliers as those values that had a standardized score lower than -3.29 or higher than 3.29 (Tabachnick et al., 2013). Per variable, zero to ten outliers were determined. 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. When this led to differences in statistical significance of results (this was the case for 3 effects, see footnote under Table 3), we additionally reported the effect of the analyses with outliers. In none of the cases however, running the analyses with or without outliers affected the conclusions that can be drawn from the outcomes.

Results

Teachers' judgments of students' monitoring skills (RQ1)

Using our framework, we investigated the degree to which teachers' judgments regarding students' monitoring skills were accurate (RQ1). First, we investigated to what degree teachers *think* that students accurately monitored their own performance (RQ1a; SMA-TJ).

Table 2 Means and standard deviations of all measures of the framework of teachers' judgments of students' monitoring skills

	n	*M*	(*SD*)
Teachers' judgment of students' performance (TJSP)[a]	406	14.60	(5.60)
Teachers' judgment of students' monitoring judgment (TJSMJ)[a]	397	14.63	(4.65)
Student performance (SP)[a]	401	10.94	(6.04)
Students' monitoring judgment (SMJ)[a]	387	12.85	(5.04)
Students' monitoring accuracy—bias (SMA)[b]	387	1.61	(5.34)
Students' monitoring accuracy—absolute deviation[a] (SMA)	387	4.46	(3.34)
Teachers' judgment accuracy—bias[b] (TJA)	401	3.66	(5.70)
Teachers' judgment accuracy—absolute deviation[a] (TJA)	401	5.27	(4.26)
Student monitoring accuracy according to teacher judgments—bias[b] (SMA-TJ)	397	-0.003[c]	(5.08)
Student monitoring accuracy according to teacher judgments—absolute deviation[a] (SMA-TJ)	397	4.04	(3.08)
Teachers' judgment accuracy of students' monitoring judgment—bias[b] (TJASMJ)	378	1.85	(5.52)
Teachers' judgment accuracy of students' monitoring judgment – absolute deviation[a] (TJASMJ)	378	4.51	(3.68)
Correctness of student monitoring accuracy according to teacher judgments[d] – bias (C-SMA-TJ)	378	-1.69	(6.77)
Correctness of student monitoring accuracy according to teacher judgments—absolute deviation (C-SMA-TJ)[a]	378	3.44	(2.99)

[a] Range: 0 to + 24

[b] Range: –24 to + 24

[c] Does not significantly differ from zero

[d] Range: –48 to + 48

On average, teachers' judgments of students' monitoring judgments (TJSMJ) deviated 4.04 points from teachers' judgments of students' performance (TJSP), meaning that teachers may think that students' misjudged their own understanding by about four points or (4.04/24 =) 16.83% (i.e., SMA-TJ-absolute). Student monitoring accuracy according to teacher judgments did not show a tendency towards overestimation or underestimation (SMA-TJ-bias).

Second, we explored the degree to which the teachers' judgments of students' monitoring judgments were accurate (RQ1b; TJASMJ). Teachers' judgment accuracy of student monitoring judgment (TJASMJ-absolute) was 4.51 points (4.52/24 = 18.79%). In other words, the teachers' judgments of students' monitoring judgments (TJSMJ) deviated 4.51 points from students' actual monitoring judgments (SMJ). On average, there was a slight tendency for teachers to overestimate students' monitoring judgments, as the bias score was + 1.85 (TJASMJ-bias).

Finally, we explored the extent to which student monitoring accuracy according to teacher judgments was in line with the student's actual monitoring accuracy (RQ1c; C-SMA-TJ). Teachers' judgments of students' judgment accuracy deviated 3.44 points (3.44/24 = 14.33%) from the actual accuracy of students' own monitoring judgments (C-SMA-TJ-absolute). In addition, the bias score for the teachers' correctness of student monitoring accuracy according to teacher judgments (C-SMA-TJ-bias) was -1.69, meaning that teachers on average

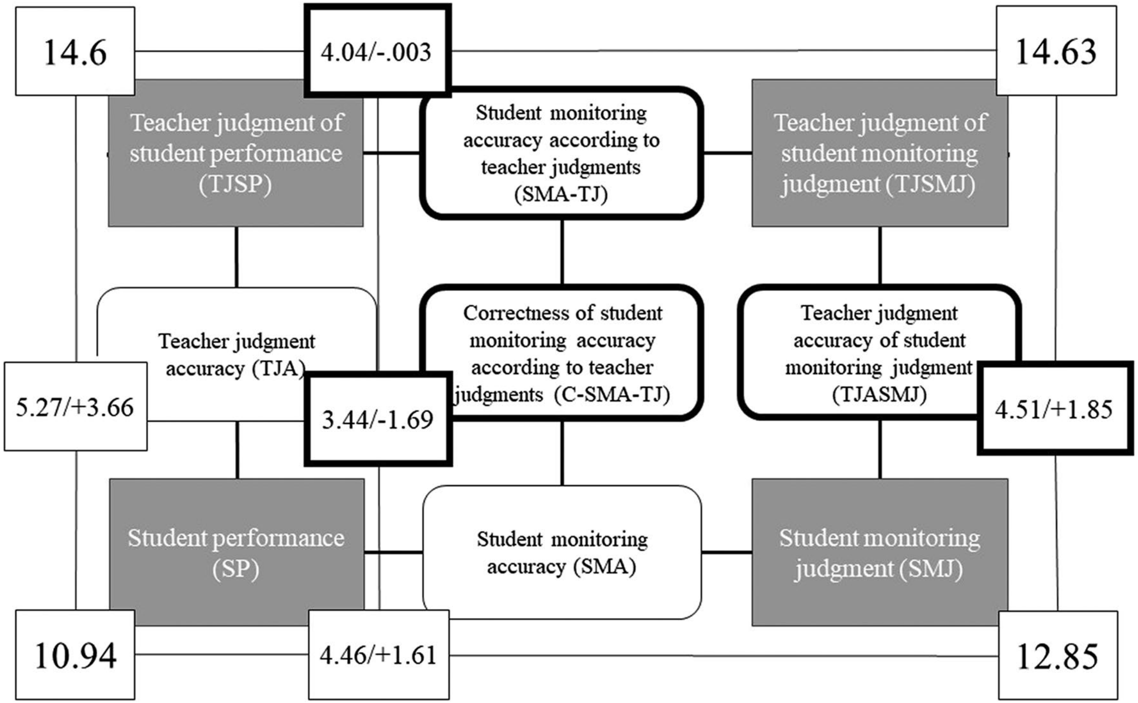

14.6
4.04/-.003
14.63
Teacher judgment of student performance (TJSP)
Student monitoring accuracy according to teacher judgments (SMA-TJ)
Teacher judgment of student monitoring judgment (TJSMJ)
Teacher judgment accuracy (TJA)
Correctness of student monitoring accuracy according to teacher judgments (C-SMA-TJ)
Teacher judgment accuracy of student monitoring judgment (TJASMJ)
5.27/+3.66
3.44/-1.69
4.51/+1.85
Student performance (SP)
Student monitoring accuracy (SMA)
Student monitoring judgment (SMJ)
10.94
4.46/+1.61
12.85

Fig. 3 Outcomes of the current study presented in the framework of teachers' judgments of students' monitoring skills. *Note.* First number represents the absolute accuracy score, second number the bias score

Table 3 Cue-diagnosticity and teachers' cue-utilization for students' performance and students' monitoring judgments for the total sample and groups with low and high accuracy

	Cue	Teachers judgments of students' performance				Teacher judgments of students' monitoring judgments			
		Diagnosticity for students' test score	Teachers' cue-utilization			Diagnosticity for students' monitoring judgment	Teachers' cue-utilization		
			Total sample	High TJA	Low TJA		Total sample	High TJASMJ	Low TJASMJ
Performance cues	Question marks in students' diagram	**-0.54****	**-0.57****	**-0.62****	-0.25*	**-0.54****	-0.31**	**-0.44****	0.00
	Omission errors in students' diagram	**-0.74****	**-0.68****	**-0.88****	-0.20	**-0.56****	-0.37**	**-0.63****	-0.08
	Commission errors in students' diagram	-0.37**	-0.19**	-0.37**	0.06	-0.19**	-0.05	-0.23	0.06
	No. of correct facts in diagram	0.14*	0.11*	0.24*	-0.07	0.11*	-0.01	0.06	-0.09
	No. of correct elements in diagram	**0.72****	**0.62****	**0.85****	0.15	**0.53****	0.30**	**0.56****	0.07
	No. of correct relations in diagram	**0.71****	**0.61****	**0.85****	0.14	**0.51****	0.31**	**0.60****	0.04
	Average no. of words in diagram	**0.46****	**0.48****	**0.66****	0.02	**0.45****	0.25**	0.34*	-0.09
	Time needed to finish diagram	0.07	0.14**	0.11	0.04	0.14*	0.07	0.25* [c]	0.00

Table 3 (continued)

	Cue	Teachers judgments of students' performance				Teacher judgments of students' monitoring judgments			
		Diagnosticity for students' test score	Teachers' cue-utilization			Diagnosticity for students' monitoring judgment	Teachers' cue-utilization		
			Total sample	High TJA	Low TJA		Total sample	High TJASMJ	Low TJASMJ
Student cues	General effort in class	0.09	0.06	0.24	-0.07	0.12*	0.03	-0.20	-0.19
	Conscientiousness	0.05	0.03	0.07	-0.02	0.06	0.04	-0.20	0.00
	General reading comprehension level	0.32**	0.19*	0.36**	-0.04	0.19**	0.17*	0.19	0.24*
	General ability to reproduce facts	0.21**	0.20**	0.34**	-0.15	0.11*	0.10	0.11	-0.11
	Interest in text topic	0.21**	0.17*	**0.42****	-0.07	0.22*	0.12*	0.31*	-0.02
	Extraversion	-0.05	0.02	-0.11	-0.09	0.11	0.08	0.10	-0.02
	Self-efficacy subject teacher	0.14*	0.25**	0.26*	-0.03	0.22**	0.31**	**0.46****	0.19
	Gender	-0.10	-0.04	-0.17	0.11	0.07	0.04	0.25*	-0.09
	Learning problems	-0.09	-0.09	-0.27*	0.00	-0.09	-0.11	-0.12	-0.17
	General school performance	0.24*	0.27**	**0.45***	0.21	0.07	0.33**	**0.44****	0.38*
	Nationality	0.16**	0.07[a]	0.12[b]	0.19*	0.03	0.05	0.08	0.13
	Student IQ	0.30**	0.20**	0.33**	-0.02	0.20**	0.23**	**0.47****	0.02
	Self-concept reading comprehension	0.22**	0.20**	0.30**	-0.02	0.29**	0.16**	**0.54****	0.03
Task cue	No. of difficult words in the texts	-0.18**	-0.18*	**-0.43****	0.04	-0.15*	-0.13*	-0.30*	0.13

All values are standardized regression coefficients from multilevel regression analyses. Cues that have a coefficient of ≥ 0.40 and are significant are considered diagnostic/used (Evans, 1996) and are printed in bold. TJA = Teacher judgment accuracy of student performance; TJASMJ = Teacher judgment accuracy of students' monitoring judgment. Effect with outliers still included: [a]$\beta = 0.17$**; [b]$\beta = 0.18$*; [c]$\beta = 0.18$

underestimated their students' monitoring accuracy. That is, teachers thought that the student underestimated their understanding more than the student actually did.

Teachers' absolute judgments accuracy of students' performance (TJA; Table 2 and Fig. 3) was 5.27 and their judgments of students' performance (TJP) thus deviated (5.27/24=) 21.96% from students' actual performance (SP). They showed a tendency to overestimate students' performance as the bias score for teachers' judgment accuracy was + 3.37 (TJA-bias).

Students' monitoring judgments about their own performance deviated on average 4.46 points (4.46/24 = 18.58%) from students' actual performance, with a tendency to overestimate their performance (SMA-bias = + 1.61).

Cue-diagnosticity (RQ2)

Both for students' performance and for students' monitoring judgments, the number of question marks, omission errors, correct elements, correct relations, and the average number of words in students' diagrams showed moderate (0.40—0.59) to high (0.60—0.79) correlations with students' performance and with students' monitoring judgments respectively and are thus diagnostic for both outcomes (Table 3). This also means that, contrary to what we expected, students may have used cues that are highly diagnostic, given that the cue-diagnosticity for student monitoring judgments is an indication of students' cue-utilization for their own monitoring judgments of their performance. The cues explain 71% ($R^2=0.71$, $SE=0.04$, $p\leq 0.00$) of the variance of students' performance and 47% ($R^2=0.47$, $SE=0.07$, $p\leq 0.00$) of the variance of students' monitoring judgments (both indicating a large effect size of respectively $f^2=2.45$ and $f^2=0.89$ (Cohen, 1992).

Teachers' cue-utilization for judging students' performance (RQ3)

All cues together explained 57% ($R^2=0.57$, $SE=0.07$, $p\leq 0.00$) of the variance in teachers' judgments of students' performance (large effect size: $f^2=1.33$). When judging students' performance, teachers seemed to use those cues that had high diagnostic values (question marks, omission errors, correct relations, and elements and the average number of words in students' diagrams; Table 3) and not, opposed to our expectation, also low diagnostic values (e.g., IQ and effort in class; RQ3a).

We compared teachers' cue-utilization of the highly diagnostic cues for accurate and inaccurate (in terms of absolute accuracy) judgments (RQ3b). As expected, cue-utilization appeared significantly higher for accurate judgments than for inaccurate judgments for all diagnostic cues (diagnosticity values of $\geq$0.40; all diagram cues depicted in Table 4). In other words, teachers used the highly diagnostic cues to a greater extent for accurate than for inaccurate judgments.

Contrary to what we expected, the cue-utilization values of students' interest in the text topic and the number of difficult words in the text – which were not highly diagnostic –were significantly higher for the accurate judgments (absolute accuracy) compared to the inaccurate judgments. Finally, even though teachers' cue-utilization value for students' *general school performance* was higher than 0.40 and significant for accurate judgments which was not the case for the inaccurate judgments, the difference in cue-utilization for this cue between

Table 4 Differences in cue-utilization per cue for accurate and inaccurate teachers' judgments of students' performance and accurate and inaccurate teachers' judgments of students' monitoring judgments (in Terms of Absolute Accuracy)

		B	SE	*p*	R^2
Teachers' judgments of students' performance					
Performance cues	No. of question marks in diagram	0.98	0.26	0.000	0.44
	No. of omissions in the diagram	1.38	0.23	0.000	0.62
	No. of correct elements in diagram	-1.30	0.28	0.000	0.58
	No. of correct relations in diagram	-1.43	0.24	0.000	0.58
	Average no. of words in diagram	-2.73	0.58	0.000	0.42
Student cues	Students' interest in the text topic	-5.25	1.69	0.002	0.27
	General school performance	-2.19	1.28	0.087	0.40
Task cue	No. of difficult words in the text	0.85	0.19	0.000	0.29
Teachers' judgments of students' monitoring judgments					
Performance cues	No. of question marks in diagrams	1.15	0.25	0.000	0.13
	No. of omissions	0.97	0.28	0.001	0.18
	No. of correct elements in diagrams	-0.79	0.32	0.013	0.14
	No. of correct relations in diagrams	-0.90	0.26	0.001	0.16
Student cues	Students' self-efficacy subject teacher	-1.98	1.27	0.119	0.11
	Students' IQ	-0.98	0.21	0.018	0.13
	Students' self-concept	-2.82	0.86	0.001	0.13
	General school performance	-0.45	1.18	0.670	0.26

We only analyzed the interaction effects for RQ3b and RQ4b for those cues that had significant correlations of ≥ (-)0.40 for cue-use in the whole sample, accurate sample, or inaccurate sample

the accurate and inaccurate judgments was not significant. Effect sizes for all cues were substantial (Table 4; Cohen, 1988).

Teachers' cue-utilization for judging students' monitoring judgments (RQ4)

All cues together explained 42% of the variance in teachers' judgments of students' monitoring judgments (large effect size: $f^2 = 0.72$). When judging students' monitoring judgments, there were no cues that may have been used to a moderate or high extent for students' monitoring judgments as all values are below 0.40 (Table 3; RQ4a).

As expected, teachers' cue-utilization values for all diagnostic performance cues were significantly higher for the accurate judgments compared to the inaccurate judgments (absolute accuracy; Table 4; RQ4b; all moderate effect sizes; Cohen, 1988). In other words, teachers seemed to have used these highly diagnostic cues to a greater extent for accurate judgments than for inaccurate judgments. Yet, opposed to what we expected, teachers' cue-utilization values were also significantly higher for accurate judgments regarding some of the *non-diagnostic cues* (i.e., students' IQ and students' self-concept for reading comprehension; Table 4; moderate effect sizes; Cohen, 1988) compared to inaccurate judgments.

In other words, teachers also used some non-diagnostic cues to a greater extent for accurate judgments than for inaccurate judgments.

Although teachers' cue-utilization value for students' *self-efficacy* regarding the subject the teacher teaches and *general school performance* was higher than 0.40 and significant for accurate judgments, which was not the case for inaccurate judgments, the difference in cue-utilization for this cue between the accurate and inaccurate judgments was not significant.

Discussion

In this study, we presented a novel conceptual and methodological framework for studying teachers' judgments of students' monitoring skills as integrated with teachers' judgments of students' performance and students' judgments of their own performance. To be able to help students improve their monitoring skills effectively and efficiently, teachers should have an accurate idea of how well students can monitor their learning. We explored this crucial teacher skill in secondary education teachers. In addition, using the framework enabled us to explore explanations for differences in judgment accuracy by focusing on the diagnosticity and utilization of cues for teachers' judgments of students' monitoring judgments.

Teachers' judgments of students' monitoring skills (RQ1)

Categoric approaches used in previous research of Fleury-Roy and Bouffard (2006) and Jamain (2019) –asking teachers to categorize students' into those who can accurately monitor their own performance, those who overestimate their own performance and those who underestimate their performance– do not match the actual classroom situation (see Introduction) and do not indicate the *degree* of judgment (in)accuracy. Our framework does enable the investigation of the *degree* to which teachers' are able to judge students' monitoring skills, using several measures. First, the teacher's judgment of the student's monitoring accuracy (SMA-TJ) indicates the degree to which the teacher may *think* the student accurately monitors their own performance. In our sample, teachers may have thought that students' monitoring judgments about their own understanding were inaccurate by about 17% (no tendency for over- or underestimation). With the teacher judgment accuracy of the student's monitoring judgment (TJASMJ), one can check whether the teacher's idea of a student's monitoring judgment is correct. In our sample, teachers misjudged students' monitoring judgments by about 19% (slight tendency to overestimate). Finally, the correctness of students' monitoring accuracy according to teachers' judgments (C-SMA-TJ) indicates the extent to which the teacher's idea of a student's monitoring accuracy is in line with the student's actual monitoring accuracy. Teachers in our sample misjudged students' monitoring accuracy with about 14% (slight tendency to underestimate students' monitoring accuracy). Although the categoric approach of Fleury-Roy and Bouffard (2006) indicates that monitoring accuracy of about 40–60% of the students is judged inaccurately by their teachers, our findings show that the *degree* of misjudgment sketches a somewhat more positive picture: teachers' judgments of students' judgment accuracy by 14.33% from the actual accuracy of students' own monitoring judgments.

The task that was used (completing diagrams about the causal relations in the text) was not necessarily familiar to the students and teachers (they did practice the task, but they had not used this task before in their previous lessons). Therefore, students' and teachers' judgment accuracy that was found in the current study may differ from studies using tasks that are familiar to students and teachers.

Cue-diagnosticity (RQ2)

Contrary to what we expected, cues that were diagnostic for students' performance, were also diagnostic for students' monitoring judgments (i.e., the number of question marks, omission errors, correct elements, correct relations, and the average number of words in students' diagrams). We expected that the cues that students normally tend to use, which are typically non-diagnostic task or student cues (Bol et al., 2010; Hacker et al., 2008; Thiede et al., 2010), would be diagnostic for students' monitoring judgments (See "Cue-Diagnosticity for Judgments of Students' Monitoring" section). Our finding could also mean that, contrary to what is reported in previous research (Bol et al., 2010; Hacker et al., 2008; Thiede et al., 2010), students may have mostly used diagnostic cues when judging their own performance (as the cue diagnosticity for teachers' judgments of students' monitoring judgments indicates the cue-utilization for students' judgments of their own performance). As opposed to the previous studies (Bol et al., 2010; Hacker et al., 2008; Thiede et al., 2010) that asked students to predict their test scores, we asked students to estimate their test scores after having made the test (post dictions). Having information about and having made the actual test, may have affected students' cue-utilization. Whether this is the case and why would be an interesting topic for future research. Moreover, it could be that there are other cues, that we did not measure, that are more used by students (and thus more diagnostic for their monitoring judgments) than the ones measured in the present study. Future research could for example test whether other task cues than the one we included (i.e., number of difficult words in the text) and experience-based cues such as reading fluency, are used by students in judging their performance and would thus be diagnostic for teachers when judging students' monitoring judgments (Thiede et al., 2010).

Furthermore, a recent study has indicated that there are individual differences in the degree to which particular cues are diagnostic for students' performance (Van de Pol et al., 2020). Therefore, it may be worthwhile to investigate whether there are individual differences in the degree to which cues are diagnostic for students' monitoring judgments (and thus whether there are individual differences in students' cue-utilization when judging their own performance). Findings of such future studies could help to further improve interventions for teachers to improve their judgments of students' monitoring judgments.

Teachers' cue-utilization for judging students' performance (RQ3) and students' monitoring judgments (RQ4)

Overall, teachers used those cues that were diagnostic for judging students' *performance* (i.e., question marks, omission errors, correct relations, and elements and the average number of words in students' diagrams) and not – contrary to what we expected based on a previous study with similar materials – some cues with low diagnostic values. In real educational settings, teachers do not always have a practice task (here: students' completed diagrams) available that corresponds so closely with what is tested or monitored as was the case in this and our other studies. Future research could therefore investigate teachers' cue-utilization for situations in which teachers do not have practice tasks available that correspond so closely to what is tested (e.g., a summary).

When judging students' *monitoring* judgments, there were no clear indications of teachers' using particular cues in the overall sample. The latter was contrary to what we expected based on previous research, which found that teachers' used cues regarding students' academic performance, students' behavior in class, and students' self-concept (Carr & Kurtz-Costes, 1994; Dignath & Sprenger, 2020; Friedrich et al., 2013).

For accurate judgments of students' *performance*, teachers seemed to use the diagnostic cues to a greater extent than for inaccurate judgments, as expected. Opposed to what we expected, teachers also seemed to use two low diagnostic cues (students' interest in the text topic and difficult words in the text) to a greater extent for accurate judgments than for inaccurate judgments.

When judging students' *monitoring judgments*, teachers seemed to make more use of diagnostic cues (except average number of words in students' diagrams) for accurate than for inaccurate judgments, as expected. Opposed to what we expected, teachers also seemed to have used two student cues with lower diagnostic values (IQ and students' self-concept regarding reading comprehension) to a greater extent for the accurate versus inaccurate judgments. Ignoring those non-diagnostic cues may further improve teachers' judgments of students' monitoring judgments. We do like to note here that we used a shortened version of the Raven Progressive Matrices to measure IQ. Although this short version has shown to perform similar to the longer versions, the internal consistency in our sample was moderate. Therefore, we should interpret our finding regarding students' IQ with caution.

Limitations and future research

When interpreting the findings, some limitations need to be taken into account. First, we cannot draw causal conclusions regarding the effects of teachers' cue-utilization on their judgment accuracy of students' monitoring skills, given the correlational design of the current study. It is unlikely that the teachers' judgment accuracy of students' monitoring skills causes their cue-utilization. Yet, it is possible that teachers used cues that are related to the cues we have measured. Future research could therefore explore other cues that were not measured in this study. In addition, to examine the causality, future research could manipulate for example what cues teachers have available (e.g., only diagnostic, only non-diagnostic, or both diagnostic and non-diagnostic) to determine whether it is actually the teachers' cue-utilization that determines their judgment accuracy.

Furthermore, and related to the previous point, we used correlations between *actual* cue values (based on students' work or student questionnaires) and teachers' judgments as an indication for cue-utilization, following previous research (e.g., Van Loon et al., 2014). Yet, teachers often do not know the actual value of a cue but instead use their perceptions of the cue (e.g., a teacher may not know the actual conscientiousness level of a student; Van de Pol et al., 2021b). Future research may explore other measures of cue-utilization (see Van de Pol et al. (2021b) for a self-report measure and Van de Pol et al. (2020) for a discussion of different measures for cue-utilization) to determine whether such measures of teachers' cue-utilization sketch a similar or different picture of their cue-utilization and the relation between their cue-utilization and judgment accuracy.

The cues we measured were based on previous research about teacher judgments of student *performance*. It is possible that teachers base their judgments of students' *monitoring* (additionally) on cues that we did not measure, such as students' general ability to reflect on their own learning, emotions, or behavior. However, the cues we used explained a substantial amount of the variance in teachers' judgments of students' monitoring judgments (42%), showing that these cues cover a large part of the cues teachers' use. Yet, future research can further explore on which cues, beyond the cues measured in the present study, teachers base their judgments of students' monitoring as there was still variance to be explained.

Furthermore, we described the teacher's judgment of the student's monitoring accuracy (SMA-TJ) as an indication of the degree to which the teacher may *think* the student accurately monitors their own performance. Yet, this assumes that teachers think that their own judgments of students' performance or of students' monitoring judgments are accurate. Although teachers make these judgments to the best of their abilities, the degree to which they are confident that their judgments are accurate may differ. If a teacher is highly unconfident about their own judgment of a student's performance, then the teachers' judgment of the student's monitoring accuracy (SMA-TJ) may not always describe the degree to which the teacher *thinks* that they are accurate. Future research could investigate this issue, for example by interviewing teachers or asking teachers to reflect on their judgments and ask them directly how accurate they think their judgments are. Related to this, it would be interesting to measure teachers' confidence in their judgments of students' performance and of students' monitoring judgment to find out to what extent teachers are aware about their accuracy (Gabriele et al., 2016). Preferably, teachers are aware about the (in)accuracy of their judgments. When being aware of their *inaccuracy* (i.e., they know that they are inaccurate) teachers will probably first seek more information and not act upon their judgments. When being aware of their *accuracy* (i.e., they know that they are accurate), they can (justly) act upon their judgments. Furthermore, it would be interesting to – instead of asking them to give a judgment of student monitoring judgments – ask teachers directly how accurate they think that students can monitor their own understanding and to compare this direct measure to the indirect measures in the current conceptual framework.

Related to this, an interesting next step for future research would be to investigate how teachers' ability to accurately monitor students' monitoring skills affects the regulation of their students' self-regulated learning. Do teachers indeed mainly or first help students who overestimate or underestimate themselves to better monitor (and regulate) their own learning? And would their help be different for students who overestimate themselves versus those who underestimate themselves?

In addition, previous research has shown that the correctness of teachers' conceptualization of students' self-regulation (as compared to a theoretical conceptualization) was related to their self-reported cue-utilization (Dignath & Sprenger, 2020). It would be interesting to investigate whether there are links between the correctness of teachers' conceptualization of students' monitoring on the one hand, and the correctness of student monitoring accuracy according to the teacher judgments, and their cue-utilization on the other hand. That is, do teachers for example use diagnostic cues to a greater extent and are their judgments more accurate when their conceptualization of students' monitoring is more correct?

Asking teachers to first make judgments of students' performance and then make judgments of students' monitoring judgments could have affected the height of the latter judgments. That is, teachers may have used their judgments of students' performance as an anchor to make the judgments of students' monitoring judgments. In the current study, the correlation between these two judgment types was 0.525 ($p = 0.000$) showing some degree of dependency. Future research could further investigate this dependency by asking teachers to make judgments of students' monitoring judgments with and without first making judgments of students' performance.

Finally, we used the conceptual and methodological framework for measures of absolute accuracy (absolute deviation and bias). Future research could explore whether and how the model can be used for relative accuracy (e.g., intra-individual gamma correlations; Van Loon et al., 2014). To compute relative accuracy one needs several measurement occasions for each judgment (e.g., student monitoring judgment, teacher judgment of student performance, and teacher judgment of student monitoring judgment) and students' performance scores of these measurement occasions. In some studies, as few as six observations are used to calculate these gamma correlations (e.g., Van Loon et al., 2014); yet, research suggests that more observations are needed for stable estimations (e.g., Gans & Robertson, 1981). In our study, we only had three measurement occasions because having six occasions would be too time consuming for students and teachers, so calculating relative accuracy was not possible.

Conclusion

The conceptual and methodological framework can give a detailed idea of teachers' ability to judge students' monitoring skills. Accurately judging students' monitoring skills is pivotal in being able to help students become better self-regulated learners. The framework can be used as a tool to further study this important topic and help to find explanations for differences in teachers' judgment accuracy by focusing on their cue-utilization, both for judgments of students' performance and of students' monitoring judgments. The results of this study point out that in order to further improve teachers' ability to accurately judge students' monitoring, it may be worthwhile to help teachers focus their attention even more towards cues that are diagnostic of students' monitoring.

Appendix

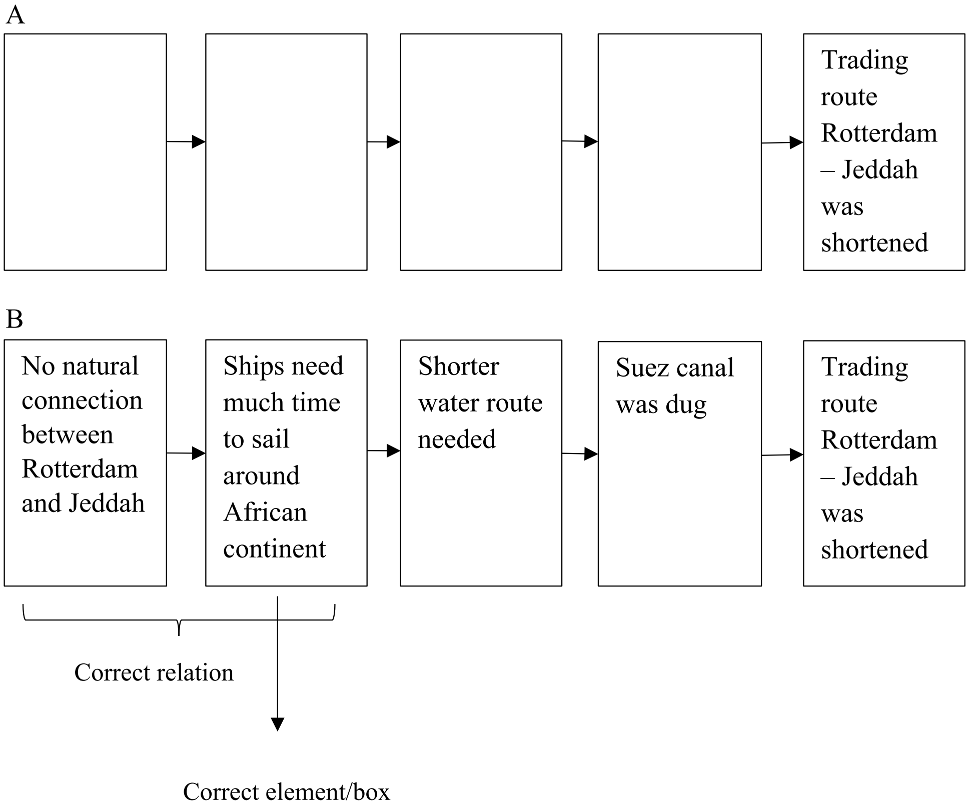

A
Trading route Rotterdam – Jeddah was shortened
B
No natural connection between Rotterdam and Jeddah
Ships need much time to sail around African continent
Shorter water route needed
Suez canal was dug
Trading route Rotterdam – Jeddah was shortened
Correct relation
Correct element/box

Fig. 4 An empty (**A**) and a correctly completed diagram (**B**) for the text 'Suez Canal

Test question for the text 'Suez Canal'

The Suez Canal

The trading route between Jeddah and Rotterdam has been significantly shortened. What are the causes for this? Mention in your answer four relations and five elements. Please be as complete as possible.

Clearly indicate the order between the different causes and consequences. You can use the words and phrases 'and', 'therefore', because of that', for that reason', 'for those two reasons', 'first', 'second', or 'this has two consequences.

Use somewhere in your answer the element: "The trading route between Jeddah and Rotterdam was shortened".

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11409-023-09349-8.

Acknowledgements We would like to thank Nynke Heegstra, Marloes Berkers, Jonne Bloem, Luca Clercx, and Susan Ravensberger who helped collecting and coding the data. We would like to thank Tamara van Gog for helping in developing the terminology in the framework.

Funding During the realization of this research the first author was funded by a Veni grant from the Dutch Research Council awarded to the first author (grant number: 451–16-012).

Data Availability Data is available upon request.

Data availability (data transparency) Data have been stored anonymously on a secure server at Utrecht University and can be made available upon request.

Code availability (software application or custom code) Not applicable.

Declarations

Ethics approval We complied with the APA ethical standards for treatment of human participants, informed consent, and data management. The study was approved by the Ethics committee of the Faculty of Behavioural Sciences of Utrecht University (file number FETC17-118).

Consent to participate Before participation, all students and parents of all students were informed about the study procedure and aims. Students were asked for active consent; parents – for students between 13 and 16 years old – were given the opportunity to decline participation of their child.

Consent for publication All authors agreed with the content and gave explicit consent to submit this study.

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

References

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. *Assessment, 19*(3), 354–369. https://doi.org/10.1177/1073191112446655

Bol, L., Riggs, R., Hacker, D. J., & Nunnery, J. (2010). The calibration accuracy of middle school students in math classes. *Journal of Research in Education, 21*(2), 81–96.

Brunswik, E. (1952). The conceptual framework of psychology. *International encyclopedia of unified science*. University of Chicago Press.

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

Carr, M., & Kurtz, B. E. (1991). Teachers' perceptions of their students' metacognition, attributions, and self-concept. *British Journal of Educational Psychology, 6*(12), 197–206. https://doi.org/10.1111/j.2044-8279.1991.tb00975.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

Cleary, T. J., & Zimmerman, B. J. (2004). Self-regulation empowerment program: A school-based program to enhance self-regulated and self-motivated cycles of student learning. *Psychology in the Schools, 41*(5), 537–550. https://doi.org/10.1002/pits.1017743

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

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

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

Destan, N., & Roebers, C. M. (2015). What are the metacognitive costs of young children's overconfidence? *Metacognition and Learning, 10*(3). https://doi.org/10.1007/s11409014-9133-z

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

Dompnier, B., Pansu, P., & Bressoux, P. (2006). An integrative model of scholastic judgments: Pupils' characteristics, class context, halo effect and internal attributions. *European Journal of Psychology of Education, 21*(2), 119–133. https://doi.org/10.1007/BF03173572

Dusek, J. B., & Joseph, G. (1983). The bases of teacher expectancies: A meta-analysis. *Journal of Educational Psychology, 75*(3), 327–346. https://doi.org/10.1037/0022-0663.75.3.327

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*(4), 228–232. https://doi.org/10.1111/j.1467-8721.2007.00509.x

Dunlosky, Mueller, M. L., & Thiede, K. W. (2016). Methodology for investigating human metamemory: Problems and pitfalls. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199336746.013.1

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current directions in psychological science, 12*(3), 83–87. https://doi.org/10.1111/1467-8721.01235

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole Publishing.

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

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

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

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

Gans, L. P., & Robertson, C. A. (1981). Distributions of Goodman and Kruskal's gamma and Spearman's rho in 2× 2 tables for small and moderate sample sizes. *Journal of the American Statistical Association, 76*(376), 942–946. https://doi.org/10.1080/01621459.1981.10477745

Glock, S., Krolak-Schwerdt, S., Klapproth, F., & Böhmer, M. (2012). Improving teachers' judgments: Accountability affects teachers' tracking decisions. *International Journal of Technology and Inclusive Education*, *1*, 89–98. http://dx.doi.org.proxy.library.uu.nl/10.1007/s10212-015-0259-4

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

Griffin, T., Mielicki, M., & Wiley, J. (2019). Improving Students' Metacomprehension Accuracy. In J. Dunlosky & K. Rawson (Eds.), *The Cambridge Handbook of Cognition and Education* (Cambridge Handbooks in Psychology, pp. 619–646). Cambridge University Press.

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning, 3*(2), 101–121. https://doi.org/10.1007/s11409-008-9021-5

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

IRRE. (1998). *Student effort: Institute for Research and Reform in Education. Research Assessment Package for Schools (RAPS) manual for elementary and middle school assessments*. Retrieved July 27, 2023, from http://www.irre.org/publications/research-assessment-package-schools-raps-manual

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, Grenoble Alpes). Retrieved July 27, 2023, from: https://www.theses.fr/2019GREAH006.pdf

Kamalski, J. (2007). *Coherence Marking, Comprehension and Persuasion on the processing and representations of discourse* (Doctoral dissertation, Netherlands Graduate School of Linguistics). Retrieved July 27, 2023, https://dspace.library.uu.nl/bitstream/handle/1874/22617/full.pdf?sequence=1&isAllowed=y

Kaufmann, E., & Athanasou, J. A. (2009). A meta-analysis of judgment achievement as defined by the lens model equation. *Swiss Journal of Psychology/schweizerische Zeitschrift Für Psychologie/revue Suisse De Psychologie, 68*(2), 99. https://doi.org/10.1024/1421-0185.68.2.99

Kostons, D., Van Gog, T., & Paas, F. (2012). Training self-assessment and task-selection skills: A cognitive approach to improving self-regulated learning. *Learning and Instruction, 22*(2), 121–132. https://doi.org/10.1016/j.learninstruc.2011.08.004

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Lin, L. M., Zabrucky, K. M., & Moore, D. (2002). Effects of text difficulty and adults' age on relative calibration of comprehension. *American Journal of Psychology, 115*(2), 187–198. https://doi.org/10.2307/1423434

Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S. A., & Harackiewicz, J. M. (2010). Measuring situational interest in academic domains. *Educational and Psychological Measurement, 70*(4), 647–671. https://doi.org/10.1177/0013164409355699

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*, 311–360. https://doi.org/10.1207/s15327574ijt0604_1

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

Moerbeek, M., & Teerenstra, S. (2015). *Power analysis of trials with multilevel data*. Chapman and Hall/CRC.

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

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). TIMSS 2007 technical report. TIMSS & PIRLS International Study Center

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.2022.103982

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

Prinz, A., Golke, S., & Wittwer, J. (2020). To what extent do situation-model-approach interventions improve relative metacomprehension accuracy? Meta-analytic insights. *Educational Psychology Review*, 1–33. https://doi.org/10.1007/s10648-020-09558-6

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

Schleinschok, K., Eitel, A., & Scheiter, K. (2017). Do drawing tasks improve monitoring and control during learning from text? *Learning and Instruction, 51*, 10–25. https://doi.org/10.1016/j.learninstruc.2017.02.002

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

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2013). *Using multivariate statistics* (Vol. 6, pp. 497–516). Boston, MA: Pearson.

Thiede, K. W., & Anderson, M. C. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28*(2), 129–160. https://doi.org/10.1016/S0361-476X(02)00011-5

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., Oswalt, S., Brendefur, J., Carney, M., & Osguthorpe, R. (2019). Teachers' judgments of student learning of mathematics. In J. Dunlosky & K. Rawson (Eds.), *The Cambridge handbook of cognition and education* (Cambridge Handbooks in Psychology, pp. 678–695). Cambridge University Press. https://doi.org/10.1017/9781108235631.027

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

Van Berkel, H., & Bax, A. (2006). *Toetsen in het hoger onderwijs* [Testing in higher education. Houten/Diegem: Bohn Stafleu Van Loghum.

Van de Pol, J., de Bruin, A. B., 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., van den Boom-Muilenburg, S. N., & van Gog, T. (2021a). Exploring the relations between teachers' cue-utilization, monitoring and regulation of students' text learning. *Metacognition and Learning, 16*(3), 769–799. https://doi.org/10.1007/s11409-021-09268-6

Van de Pol, J., van Gog, T., & Thiede, K. (2021b). The relationship between teachers' cue-utilization and their monitoring accuracy of students' text comprehension. *Teaching and Teacher Education, 107*,

Van de Pol, J., van Loon, M., van Gog, T., Braumann, S., & de Bruin, A. (2020). Mapping and drawing to improve students' and teachers' monitoring and regulation of students' learning from text: Current findings and future directions. *Educational Psychology Review, 32*, 951–977. https://doi.org/10.1007/s10648-020-09560-y

Van den Brink, W. P., & Mellenbergh, G. J. (1998). *Testleer en testconstructie* [Test theory and test construction]. Boom.

Van Loon, M. H., de Bruin, A. B., van Gog, T., van Merriënboer, J. J., & 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

Veenman, M. V., & van Cleef, D. (2019). Measuring metacognitive skills for mathematics: Students' self-reports versus on-line assessment methods. *ZDM Mathematics Education, 51*(4), 691–701. https://doi.org/10.1007/s11858-018-1006-5

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

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