



# Exploring the relations between teachers' cue-utilization, monitoring and regulation of students' text learning

Janneke van de Pol<sup>1</sup> · Selia N. van den Boom-Muilenburg<sup>1,2</sup> · Tamara van Gog<sup>1</sup>

Received: 24 August 2020 / Accepted: 17 May 2021 / Published online: 9 June 2021  
© The Author(s) 2021

## Abstract

This study investigated teachers' monitoring and regulation of students' learning from texts. According to the cue-utilization framework (Koriat, in *Journal of Experimental Psychology*, 126, 349–370, 1997), monitoring accuracy depends on how predictive the information (or cues) that teachers use to make monitoring judgments actually is for students' performance. Accurate monitoring of students' comprehension is considered a precondition for adaptive regulation of students' learning. However, these assumptions have not yet been directly investigated. We therefore examined teachers' cue-utilization and how it affects their monitoring and regulation accuracy. In a within-subjects design, 21 secondary education teachers made monitoring judgments and regulation decisions for fifteen students under three cue-availability conditions: 1) only student cues (i.e., student's name), 2) only performance cues (i.e., diagrams students completed about texts they had read), and 3) both student and performance cues (i.e., student's name and completed diagram). Teachers' absolute and relative monitoring accuracy was higher when having student cues available in addition to diagram cues. Teachers' relative regulation accuracy was higher when having only performance cues available instead of only student cues (as indicated by a direct effect). Monitoring accuracy predicted regulation accuracy and in addition to a direct effect, we also found an indirect effect of cue-availability on regulation accuracy (via monitoring accuracy). These results suggest that accurate regulation can be brought about both indirectly by having accurate monitoring judgments and directly by cue-utilization. The findings of this study can help to refine models of teacher monitoring and regulation and can be useful in designing effective interventions to promote teachers' monitoring and regulation.

**Keywords** Teacher monitoring · Teacher judgment accuracy · Teacher regulation · Adaptive regulation · Metacomprehension accuracy · Text comprehension

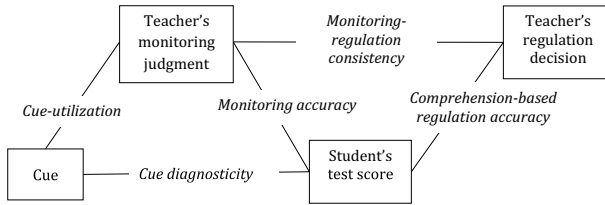
Teachers' ability to adaptively regulate students' learning is a crucial aspect of teachers' professional competence (e.g., Südkamp et al., 2012), as it has been shown to improve students' learning outcomes (Van de Pol et al., 2010). Adaptive regulation of student learning

---

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

<sup>1</sup> Department of Education, Utrecht University, PO Box 80.140, 3508 TC Utrecht, The Netherlands

<sup>2</sup> Faculty of Behavioural, Management and Social Sciences, University of Twente, Enschede, The Netherlands



Part of this figure is based on “Methodology for investigating human metamemory: Problems and pitfalls” by J. Dunlosky, M. L. Mueller, and K. W. Thiede, 2016. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford Handbook of Metamemory*. Oxford University Press.  
doi: 10.1093/oxfordhdb/9780199336746.013.1

**Fig. 1** Conceptual Model of the Relations between Cues, Monitoring Judgments, Test Scores, and Regulation Decisions

is defined in the literature as the provision of guidance or support that is adapted to a student’s current level of understanding (Hardy et al., 2019; Parsons et al., 2018; Van de Pol et al., 2010). Thus, making an accurate (i.e., adaptive) regulation decision also requires accurate *monitoring* of a student’s current level of understanding (e.g., Thiede et al., 2019).

The present study focuses on teachers’ monitoring and regulation of students’ text learning. As studying texts is an activity performed in almost every school subject, regulating students’ learning of texts is commonly required of teachers in their daily classroom practices (Artelt & Rausch, 2014). Whereas a lot of research has been conducted on (how to improve) *students’* ability to self-monitor and self-regulate their learning from texts (see Griffin et al., 2019 for a review), much less is known about (how to improve) *teachers’* monitoring and regulation of students’ text learning. Therefore, the present study investigated whether and how providing teachers with information from which students’ understanding could be inferred, would improve their monitoring and regulation accuracy.

## Monitoring and regulation accuracy

There seems to be agreement in the literature on teachers’ regulation of student learning (e.g., Thiede et al., 2019; Van de Pol et al., 2010) that accurate monitoring is a necessary but not sufficient condition for making accurate regulation decisions. Accurate monitoring is defined as accurately judging one’s own/a student’s current level of understanding and accurate regulation as accurately deciding on what subsequent study activity the learner should engage in. According to the cue-utilization framework (Koriat, 1997), to arrive at accurate monitoring judgments, teachers need to have and use bits of information (also known as ‘cues’) that are predictive or *diagnostic* of students’ comprehension (Dunlosky et al., 2016; Koriat, 1997; Thiede et al., 2015; also see Fig. 1 for a conceptual model). Generally, cues that pertain to students’ performance such as performance on a practice task are more diagnostic than cues that pertain to student characteristics (e.g., students’ gender) or that pertain to the task (e.g., text length) (Van de Pol et al., 2019; Oudman et al., 2018; Thiede et al., 2015).

In the literature, regulation accuracy is often operationalized as the extent to which teachers’ (or students’ own) judgments of what the student should do next, are aligned with their monitoring judgments of the student’s understanding (e.g., Engelen et al., 2018; Van Loon et al., 2014), regardless of the accuracy of those monitoring judgments.

In other words, this measure gives an indication of whether teachers (or students) use their monitoring judgments in making regulation decisions. This has been referred to as a *monitoring-based measure of regulation accuracy* (Van de Pol et al., 2020; Van Loon et al., 2014; also see Fig. 1). For sake of clarity, we refer to this as monitoring-regulation consistency, as it pertains to the degree to which teachers' regulation judgments are in line with their monitoring judgments. Note though, that defining regulation accuracy as the consistency between monitoring and regulation judgments can lead to a situation in which high regulation accuracy does not necessarily mean that the regulation decision is adapted to the student's actual needs. That is, the teacher or student may base their regulation decision on their monitoring judgment, but if that monitoring judgment is an overestimation or underestimation of the student's performance, the regulation decision does not match the student's needs.

That problem can be circumvented by defining regulation accuracy as the extent to which teachers' (or students') judgments of what the student should do next, are aligned with the student's actual understanding/performance. This has been referred to as a *comprehension-based measure of regulation accuracy* (Van de Pol et al., 2020; also see Fig. 1), and provides a *direct* measure of whether the teachers' regulation decision is in line with a student's needs. Thus, both monitoring-regulation consistency and comprehension-based regulation accuracy provide relevant, but different information (except in a situation in which monitoring judgments are perfectly accurate, in which case both measures have the same outcome). That is, monitoring-regulation consistency provides information about the extent to which monitoring and regulation judgments are aligned and thus indicate to what extent a monitoring judgment has been used to make a regulation judgment (in which case one would/could expect that interventions to successfully increase monitoring accuracy would also affect regulation accuracy). Comprehension-based regulation accuracy provides information about how the quality of a regulation judgment; that is, the extent to which the regulation judgment matches a student's actual understanding and thus indicates how successful such a regulation judgment will probably be. Therefore, including both measures in research is important for deepening our understanding of teachers' regulation and how accurate regulation is brought about.

## Empirical research on improving teachers' monitoring and regulation accuracy

A few studies have shown that teachers' monitoring accuracy and comprehension-based regulation accuracy is low, but that providing access to diagnostic cues can promote monitoring and/or regulation accuracy. As in research on students' self-monitoring and self-regulation, studies on teachers' monitoring and regulation of student learning have defined and measured monitoring accuracy in relative and absolute terms. Relative accuracy refers to the degree to which a teacher can indicate which items (e.g., word-pairs, texts, problem-solving tasks) a student has learned better relative to other items (monitoring) or which items need more restudy than others (regulation). Absolute accuracy indicates how accurately teachers know how well students have learned each item in itself (monitoring) or how much each item needs to be restudied (regulation).

Van de Pol et al. (2019) investigated how the availability of cues affected teachers' relative monitoring accuracy and relative comprehension-based regulation accuracy of students' text learning. The availability of diagnostic performance cues (i.e., cues predictive of

students' understanding, which teachers can derive from a student product made available to them, such as students' prior performance on a diagramming task about the text<sup>1</sup>) and non-diagnostic student cues (i.e., cues teachers have when knowing a student's name, such as their classroom behaviour, motivation, or socio-economic status) was manipulated. That is, teachers had access to both student and performance cues (i.e., they saw students' names and the diagrams students completed about the texts they had read) or only student cues (i.e., only seeing students' names). In both conditions, task cues were available (i.e., cues derived from the task-materials such as task-difficulty; cf. Cooksey et al., 2007; Oudman et al., 2018; Webb, 2015). Teachers' relative monitoring accuracy was low both when having only student cues available and when having student cues and performance cues available. In addition, relative monitoring accuracy was not significantly higher in the latter condition compared to the first. Teachers' relative comprehension-based regulation accuracy, however, did improve when having both performance cues and student cues available instead of only student cues. That is, teachers were better able to match restudy decisions to students' actual text comprehension when student and performance cues (e.g., diagrams) were available compared to student cues alone. As expected, teachers' regulation decisions about whether or not the student should restudy each text were strongly in line with their monitoring judgments with no differences between conditions (i.e., high monitoring-regulation consistency; between -0.68 to -0.74). Summarizing, having diagrams available helped teachers to get insight into students' needs for further study (i.e., comprehension-based regulation accuracy), but did not affect the alignment of monitoring judgments and study-needs (i.e., monitoring-regulation consistency).

Teachers in the studies of Thiede et al. (2015, 2018) also showed low levels of relative monitoring accuracy of students' math comprehension, but their absolute monitoring accuracy was somewhat higher (cf. Thiede et al., 2018). When the teachers participated in a professional development program aimed at implementing student-centered instruction in the classroom – which helped them to gather diagnostic performance cues about their students' mathematical understanding (e.g., from students' performance on practice problems performed in class) – both their relative and absolute monitoring accuracy improved.

Although accurate monitoring is seen as an important condition for accurate comprehension-based regulation, the *relation* between monitoring accuracy and regulation accuracy has – to the best of our knowledge – not been tested before. Some studies have focused on effects of cue-availability conditions on monitoring-regulation consistency (e.g., Engelen et al., 2018; gamma = -0.60 to -0.74; Van Loon et al., 2014; gamma = -0.53 to -0.74), but this pertains to the degree to which the monitoring judgments and regulation judgments are aligned and not the alignment of the regulation judgments with students' actual understanding (i.e., comprehension-based regulation accuracy).

Furthermore, some studies have tested the effects of cue-availability conditions on teachers' monitoring accuracy and regulation accuracy separately (e.g., Van de Pol et al., 2019; Engelen et al., 2018), not in conjunction. Yet, the fact that teachers' regulation decisions became more in line with students' actual understanding (i.e., comprehension-based regulation accuracy) when having access to more diagnostic cues whereas monitoring accuracy did not (Van de Pol et al., 2019), indicates that the relation between monitoring and comprehension-based regulation accuracy may not be as straightforward as has been assumed (in the study of

---

<sup>1</sup> This diagramming task stems from Van Loon et al. (2014). In this task, students complete a pre-structured diagram about the causal relations in each text after having read all texts ( $N=6$ ) and without seeing the texts.

Van de Pol et al. (2019), effects of conditions on monitoring accuracy and comprehension-based regulation accuracy were presented separately). Although no previous research has directly tested the relation between monitoring accuracy and regulation accuracy, Van de Pol et al. (2014) investigated the relation between teachers' monitoring *actions* and the extent to which teachers' regulatory actions were in line with students' actual understanding (i.e., comprehension-based regulation accuracy). Teachers who participated in a professional development program that aimed at increasing their comprehension-based regulation accuracy, asked more diagnostic questions in class that provided them with information about students' comprehension and regulated their students' learning more accurately compared to teachers who did not participate in this program. In addition, these diagnostic questions mostly preceded teachers' regulatory actions. Yet, the actual *accuracy* of these teachers' monitoring actions was not measured, nor was the relation between monitoring accuracy and comprehension-based regulation accuracy analyzed statistically.

In addition, we know of only two studies that investigated the relation between teachers' monitoring accuracy and students' learning, suggesting that this improvement in learning was related to improved regulation of the teachers. That is, Thiede et al. (2015, 2018) have shown that when teachers became more accurate at monitoring students' (mathematics) comprehension by participating in a professional development program, students performed better. Yet, teachers' comprehension-based regulation accuracy (i.e., the relation between teachers' regulation decisions and students' actual understanding) was not measured in these studies.

## Further improving teachers' monitoring and regulation accuracy

The studies discussed in the previous section show that making sure that teachers can derive and use diagnostic cues, affects their monitoring and comprehension-based regulation accuracy. However, there was room for further improvement. It can be hypothesized that, next to providing information from which diagnostic cues can be derived, *taking away* access to non-diagnostic student cues, for example by blinding students' work, may further improve teachers' monitoring and comprehension-based regulation accuracy. In other disciplines such as social psychology, biomedical science, and forensic science, blinding has been shown to be a powerful solution to mitigate biases in judgments, as it takes away many non-diagnostic cues (Robertson & Kesselheim, 2016). As stated by Robertson and Kesselheim "The classic icon of Lady Justice wearing a blindfold symbolizes the paradoxical insight that less information can sometimes produce better decisions" (p. 3). In the field of education, some initial studies suggest that blinding can improve teachers' monitoring accuracy. Oudman et al. (2018) showed that primary education teachers' monitoring of students' mathematical comprehension was more accurate when students' work on a practice task was blinded than when it was not blinded (i.e., when teachers only had student cues available or student and performance cues). Analysis of think aloud data, recorded while teachers monitored students' math comprehension, showed that when students' work was blinded, teachers made less use of student cues and task cues, which are both known to have low diagnosticity levels. Surprisingly, although teachers did not have access to student cues, they still used some. That is, they appeared to fabricate student cues: even though they did not know the identity of the student, they tried to guess characteristics of students based on the work they saw (e.g., their gender or motivation) when making monitoring judgments. In line with Oudman et al.'s findings, Engelen et al. (2018) found somewhat

more accurate teacher monitoring of students' text comprehension when primary education teachers saw blinded student work (i.e., summaries, also see Thiede & Anderson, 2003) compared to seeing work of students they knew or when only knowing a student's name (note that a direct statistical comparison between these conditions could not be made because they were distributed across two experiments).

## The current study

We investigated to what extent the availability of different types of cues affects secondary education teachers' monitoring accuracy *and* comprehension-based regulation accuracy when judging students' text comprehension. That is, we investigate to what extent cue-availability affects teachers' ability to accurately judge students' understanding (i.e., monitoring accuracy) and to make regulation decisions that are in line with students' understanding (i.e., comprehension-based regulation accuracy). In addition, we report teachers' monitoring-regulation consistency. We extend previous research by including both monitoring-regulation consistency and comprehension-based regulation accuracy and by investigating how accurate comprehension-based regulation is brought about and to what extent accurate monitoring is a necessary first step for arriving at accurate regulation.

In some conditions, teachers had access to causal diagrams that students completed about each text. In the current study, each teacher was asked to make monitoring and regulation judgments about 15 students in total; they made judgments about five different students under three conditions in a within-subjects design, in which the availability of cue-types was manipulated. Teachers were provided with: 1) only *student cues* (i.e., the name of the student to be judged; name-only condition), 2) both *student cues* and *performance cues* (i.e., both the name of the student and the completed diagram; name + diagram condition), and 3) only *performance cues* (i.e., a diagram the student completed about the texts (s)he had read, containing cues that are diagnostic of students' comprehension of causal relations in the text, cf. Van Loon et al., 2014; diagram-only condition). Task cues evoked by the texts were always available because teachers read the texts before making monitoring judgments. In addition to manipulating the availability of cues, we also measured teachers' actual cue-utilization to be able to explain differences between conditions.

The first research question was: to what extent does cue-availability affect teachers' monitoring accuracy (RQ1)? We describe our hypotheses in this section; an overview of the hypotheses is presented in Table 1. We hypothesized that teachers' monitoring of students' performance would be more accurate when only performance cues were available (diagram-only condition) than when student cues were available either as the only source of information (name-only condition; H1.1) or in addition to performance cues (name + diagram condition; H1.2; Van de Pol et al., 2019; Engelen et al., 2018; Oudman et al., 2018; Robertson & Kesselheim, 2016). Based on the cue-utilization framework (cf. Koriati, 1997) one would expect that teachers' monitoring would be more accurate when having access to student and performance cues (name + diagram condition) than when only having access to student cues (name-only condition). Yet, previous studies did not find this difference (Van de Pol et al., 2019; Engelen et al., 2018; Oudman et al., 2018). We therefore explored differences between these two conditions (E1.3).

To be able to explain possible differences between conditions, we also measured teachers' cue-utilization. Our second research question was: What performance cues, student cues, and task cues do teachers use when monitoring students' text comprehension

(RQ2)? Based on Oudman et al. (2018), we expected that teachers would mainly use: (a) student cues and task cues in the name-only condition, (b) student, task, and performance cues in the name + diagram condition, and (c) task, performance and, possibly, fabricated student cues in the diagram-only condition.

Finally, we focused on the relations between cue-availability, monitoring accuracy, and comprehension-based regulation accuracy. The third research question was: to what extent does the availability of cues directly and/or indirectly affect teachers' comprehension-based regulation accuracy (RQ3)? For comprehension-based regulation accuracy, we had the following hypotheses (see Table 1). Assuming that monitoring is more accurate in the diagram-only condition than in the name-only condition (i.e., H1.1) or name + diagram condition (i.e., H1.2) and that more accurate monitoring results in more accurate comprehension-based regulation (H3.1), we expect that there will be an *indirect* effect of condition on comprehension-based regulation accuracy via monitoring accuracy: Comprehension-based regulation accuracy will be higher in the diagram-only condition than in the name-only condition (H3.2a) or in the name + diagram condition (H3.2b). If there would be a difference in monitoring accuracy between the name + diagram and name-only condition, we would also expect an indirect effect on comprehension-based regulation accuracy (E3.2c).

We also investigated the *direct* effects of condition on comprehension-based regulation; teachers possibly use different (constellations of) cues for their monitoring and regulation decisions, which may result in direct effects of cue-availability on comprehension-based regulation (cf. Van de Pol et al., 2019). Regarding the direct effects, we explored differences in regulation accuracy between the diagram-only condition and the name-only condition (E3.3a) and the diagram-only and name + diagram condition (E3.3b). Based on the cue-utilization framework and the findings of Van de Pol et al. (2019), we hypothesized that teachers' regulation would be more accurate in the name + diagram condition compared to the name-only condition (H3.3c).

**Table 1** Overview of the Hypotheses and Findings of the Current Study

| Hypothesis   | Hypothesis supported?                     |  |
|--|---|--|
| RQ1: To what extent does cue-availability affect teachers' monitoring accuracy?                                |   |  |
| H1.1   | Diagram-only > name-only                  | No   |
| H1.2   | Diagram-only > name + diagram             | Absolute and relative accuracy:<br>Diagram only < name + diagram |
| E1.3 <sup>a</sup>  | Name + diagram ? name-only                | No difference  |
| RQ3: To what extent does the availability of cues (in)directly affect comprehension-based regulation accuracy? |   |  |
| H3.1   | Monitoring accuracy → regulation accuracy | Yes, for relative and absolute accuracy                          |
| Indirect effects   |   |  |
| H3.2a  | Diagram-only > name-only                  | No   |
| H3.2b  | Diagram-only > name + diagram             | Yes, for absolute accuracy                                       |
| E3.2c  | Name + diagram ? name-only                | No difference  |
| Direct effects   |   |  |
| E3.3a  | Diagram-only ? name-only                  | Relative accuracy: diagram-only > name-only                      |
| E3.3b  | Diagram-only ? name + diagram             | No difference  |
| H3.3c  | Name + diagram > name-only                | Relative accuracy: Name + diagram > name-only                    |

<sup>a</sup>An 'E' indicates an explorative question without pre-defined hypothesis



For *monitoring-regulation consistency* (i.e., the relation between teachers' monitoring and regulation judgments), we do not expect differences between conditions. We expect that if monitoring accuracy increases in one condition (indicating that their monitoring judgments are more in line with students' actual understanding), their comprehension-based regulation accuracy (indicating that their regulation judgments are more in line with students' actual understanding) will also increase. Yet, the consistency between teachers' monitoring judgments and regulation judgments (i.e., monitoring-regulation consistency) is not necessarily expected to increase based on cue-availability. This has also been shown in previous studies, which have shown that there were no differences in monitoring-regulation consistency under conditions differing in the availability of performance cues (Engelen et al., 2018; Van Loon et al., 2014). Therefore, we will only report the descriptive statistics for monitoring-regulation consistency.

## Method

### Design and participants

In a within-subjects design, teachers were asked to judge the test performance of five students (+one practice student per condition) under each of the three conditions. Teachers were provided with: 1) only student cues (i.e., the name of the student to be judged: name-only condition), 2) both student- and performance cues (i.e., both the name of the student and the diagram the student completed about the text: name+diagram condition), or 3) only performance cues (i.e., the completed diagram: diagram-only condition). For the student materials, six versions (based on a Latin Square Design) were used in which the order of the texts differed. Within each version, the order of the texts was kept the same for the different activities (e.g., making judgments, restudy selections).

### Teachers

A multilevel power analysis based on parameters from Van de Pol et al. (2019), with a power of 0.80, showed an  $N$  of 19 when  $\alpha=0.05$  and an  $N$  of 23 when  $\alpha=0.025$ . We managed to collect data from 21 teachers (14 women) who taught at nine different schools in the Netherlands in year 1 to 5 of pre-university or senior general secondary education. Teachers' age ranged from 22 to 58 years ( $M=39.01$ ,  $SD=10.77$ ). Their teaching experience ranged between 1 and 35 years ( $M=13.47$ ,  $SD=9.33$ ).<sup>2</sup>

### Students

Students ( $N=568$ ) participated in a parallel 'student study' in which they were assigned to a control condition (completing a picture matching task), a diagram completion condition (completing diagrams about the texts), or a diagram completion+instruction condition (completing diagrams about the texts and inspecting their diagrams afterwards with particular instruction without the opportunity to change their diagrams). Because of time

<sup>2</sup> Results of a recent review showed that "job experience is not associated or only weakly associated with judgment accuracy." (Urhahne & Wijnia, 2021, p.10).



constraints, teachers could not judge the comprehension of all students. Therefore, they made judgments about five students (+ 1 practice student) per condition, so 15 judgments in total (or 18 when counting the practice students, but these were not included in the analyses). Six students from each class were randomly selected for teacher judgments in the name-only condition and six students for teacher judgments in the name + diagram condition, resulting in a sample of 252 students ( $M_{\text{age}} = 15.08$ ,  $SD = 1.00$ ).<sup>3</sup> The six students in the name-only condition were students drawn from the control condition ( $n = 2$ ), diagram completion condition ( $n = 2$ ) and diagram completion + instruction condition ( $n = 2$ ), and the six students in the name + diagram condition were drawn from the diagram completion condition ( $n = 3$ ) and the diagram completion + instruction condition ( $n = 3$ ) in the 'student study'. For teachers' judgments in the diagram-only condition, data from six students in year 3 of pre-university secondary education ( $M_{\text{age}} = 15.28$ ,  $SD = 0.37$ ) from another study with the same materials was used (Van de Pol et al., 2019). Thus, the teachers did not know these students, which was done to ensure that teachers would not be able to guess from which student the work stemmed (e.g., by recognition of hand-writing).

## Materials

To better understand what the teachers did in this study, we first describe which tasks the students performed. Teachers namely judged students' understanding in the context of the specific student tasks that we used in this study.

### Expository texts

In this study, we used six expository texts that stem from Van Loon et al. (2014), on the following topics: "Sinking of metro cars", "Botox", "The Suez Canal", "Music makes smart", "Money does not bring happiness", and "Concrete constructions". For an example, see [Appendix A](#). The texts had 169.3 words on average ( $SD = 10.6$ ) and each contained four causal relations.

### Diagram completion task and test prediction

Students completed pre-structured diagrams about each text. Their worksheet contained, for each text, five textboxes that were connected by arrows. One of the text boxes was already completed; the students were instructed to complete the rest of the diagram (see [Fig. 2](#)).

### Test

Students completed a test consisting of six questions about the causal relations in the text (one question per text) (cf. Van Loon et al., 2014; for examples of questions, see [Appendix B](#)<sup>4</sup>). When taking the test, students were shown the title of the texts, accompanied by the

<sup>3</sup> One teacher judged only five students (i.e., four when not considering the practice student) in the diagram-only condition due to disturbance of the experiment, resulting in one missing value in the diagram-only condition.

<sup>4</sup> The test also contained questions about facts in the text, but this data was not used for the current study as it falls outside the scope of the paper.



**Fig. 2** Diagram for the Text “The Suez Canal” that Students Had to Complete. Adapted from “Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy”, by M. van Loon et al. (2014), *Acta Psychologica*, 121, p. 145. Reprinted with Permission

question. The question required an essay-answer and per question, student could gain four points, one per causal relation (24 points in total).

Students’ comprehension of the text’s causal relations was measured by scoring students’ essay answers. The performance on those questions was scored by assigning one point per correctly stated causal relation, resulting in a maximum of four points per text. As comprehension was emphasized, a response indicating gist understanding (i.e., the student did not mention what was literally stated in the text, but the students’ response indicated that they understood what was meant with the original text) was also scored as correct (cf. Van Loon et al., 2014). Two raters independently scored 198 essays (13.10% of all test responses) on causal relation questions. Inter-rater reliability was high (Krippendorff’s  $\alpha=0.83$ ; cf. Krippendorff, 2004). The test appeared to have a good internal consistency, ( $\alpha=0.78$ ;  $\omega=0.80$ ; Cronbach, 1951; Hayes & Coultts, 2020).

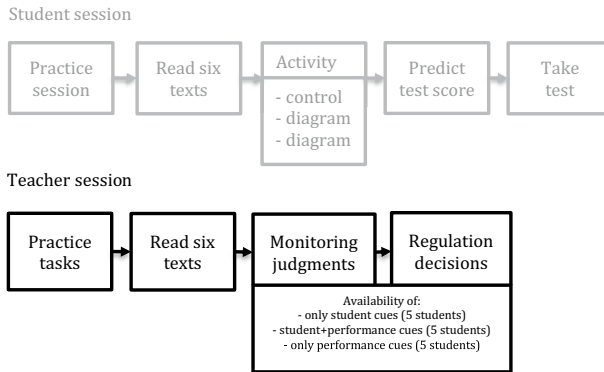
### Teachers’ monitoring judgments and regulation decisions

When making monitoring judgments, the instructions were as follows: “Have a look at this student’s diagrams (if available). Indicate below, for each text, how many questions about relations this student has answered correctly on the test.” Teachers were provided with the titles of each text and a judgment scale (0 – 4). On a next page, teachers were asked to indicate which text(s) they would advise the student to study again before making the test. Teachers were not allowed to look back at their judgments.

### Procedure

#### Students

First, a session of about 60 min took place in the classrooms of the participating teachers (see Fig. 3). This session followed the standard metacomprehension paradigm (cf. Van de Pol et al., 2020; Griffin et al., 2019). For the current study, we only used students’ diagrams and test scores; effects of the student intervention on students’ monitoring and regulation are presented elsewhere (De Bruin et al., 2017). First, students participated in a practice session. In this session, students read practice tests, answered practice test items and practiced in drawing diagrams. Then, students read all six texts in their own pace. Subsequently, students completed a pre-structured diagram about each text, without being able to see the texts. Note that this is a delayed-diagram design as the students first read all texts and only then complete the diagrams (text-order is the same for texts and diagrams within students). Furthermore, students predicted, per text, how many points they would gain on the test and which text(s) they would restudy if given a test. They were then told that they did not actually have to restudy the texts as this would influence their test score. Finally, the students took a test.



**Fig. 3** Overview Procedures of Student and Teacher Session

## Teachers

The teacher session was based on the teacher generation paradigm (cf. Van de Pol et al., 2020). Teachers saw the practice materials the students had seen and then read all six texts. Subsequently, they made monitoring judgments about students' text comprehension. Teachers made one judgments for each of the six texts; so six judgments per student. The first three students the teacher judged (i.e., one per condition, random order of conditions) were labelled practice students. Teachers did not get any feedback on their judgments; the practice students were only meant for teachers to get familiar with the materials belonging to that condition and the task of judging students' comprehension. After that, teachers judged the comprehension of fifteen students (five per condition). Finally, they indicated which text(s) they thought students should restudy before taking the test. Students and conditions were presented in a random order.

Teachers were asked to think aloud while making monitoring judgments about two randomly chosen students per condition to gain insight into teachers' cue-utilization (cf. Cooksey et al., 2007; VandeVelde et al., 2015). The researcher prompted the participants to continue thinking aloud when they stopped talking, did not give any time constraints, and did not seek clarifications or elaborations as these might interfere with the thinking processes needed for forming judgments (Ericsson & Simon, 1993). Teachers' monitoring accuracy did not differ when thinking aloud versus not thinking aloud both for relative monitoring accuracy ( $b=0.07$ ,  $SE=0.08$ ,  $p=0.40$ ) and for absolute monitoring accuracy ( $b=-0.02$ ,  $SE=0.06$ ,  $p=0.70$ ).

## Analyses

### Accuracy measures

Teachers' monitoring accuracy was measured in terms of relative accuracy and absolute accuracy. Teachers' *relative* monitoring accuracy was measured using intra-individual gamma correlations within students (Nelson, 1984; Schraw, 2009). In this measure, teachers' monitoring judgments for each of the six texts per student and students' actual test scores per text are compared to determine to what extent teachers know which texts students understand better compared to other texts (teachers judged students' understanding

for each of the six texts). So for example, judging that a student scores better on text 1 than on text 2 would add positively to a teachers' relative monitoring accuracy if the student actually scored better on text 1 than on text 2, regardless of the absolute levels of the judgments and the scores. Gamma correlations range from  $-1$  to  $+1$ . A gamma of  $+1$  indicates perfect monitoring accuracy or perfect discrimination. Following the procedure of Südkamp et al. (2008), teachers' *absolute* monitoring accuracy was defined as the mean deviation of the teachers' judgments from the students' test score on each text (cf. Dunlosky & Thiede, 2013; Schraw, 2009). The minimum deviation per judgment was 0 (i.e., perfect accuracy) and the maximum deviation was 4.

For teachers' regulation accuracy, we used two measures: monitoring-regulation consistency and comprehension-based regulation accuracy (cf. Van de Pol et al., 2020). For both of these measures, we calculated relative and absolute accuracy. Monitoring-regulation consistency refers to the extent to which teachers' regulation judgments are related to their monitoring judgments. For *relative* monitoring-regulation consistency, we used intra-individual gamma correlations within students, relating the teacher's restudy decision per text (0=no, 1=yes) to teachers' monitoring judgments per text. Teachers have high relative monitoring-regulation consistency when they, for example, indicate that a student needs to restudy text 1 but not text 2 when they gave a lower judgment for text 1 than for text 2. Therefore, values closer to  $-1$  indicate higher relative monitoring-regulation consistency. For *absolute* monitoring-regulation consistency, we used the approach of Baars et al. (2014). This regulation accuracy measure was based on each possible combination of a teacher's monitoring judgment per text and the teacher's restudy decision per text (see Table 2). Low judgments combined with decisions to restudy a text or high judgments combined with decisions to not restudy a text resulted in high absolute monitoring-regulation consistency whereas low judgments combined with decisions to not restudy a text or high judgments combined with decisions to restudy a text resulted in low absolute monitoring-regulation consistency. Absolute monitoring-regulation consistency ranged from 0 (lowest regulation accuracy) to 1 (perfect regulation accuracy) but we reversed the coding of this variable (0=perfect regulation accuracy, 1=lowest regulation accuracy) to align the interpretation of the values of this variable with that of absolute monitoring accuracy.

Comprehension-based regulation accuracy refers to the extent to which teachers' regulation judgments are related to students' test scores. For *relative* comprehension-based regulation accuracy, we used intra-individual gamma correlations, relating the teacher's restudy decision per text (0=no, 1=yes) to students' scores per text. So for example judging that a student should restudy text 1 and not text 2 would add positively to a teachers' relative comprehension-based regulation accuracy if the student actually scored better on text 2 than on text 1, regardless of the absolute scores. Gamma correlations range from  $-1$  to  $+1$  with values closer to  $-1$  indicating higher accuracy. For *absolute* comprehension-based regulation accuracy we used a similar approach as for absolute monitoring-regulation consistency (cf. Baars et al., 2014). However, for absolute comprehension-based regulation accuracy, we compared teachers' *restudy decisions* (rather than monitoring judgments) to students' test scores (see Table 2). Absolute comprehension-based regulation accuracy ranged from 0 (lowest regulation accuracy) to 1 (perfect regulation accuracy) but we reversed the coding of this variable (0=perfect regulation accuracy, 1=lowest regulation accuracy) to align the interpretation of the values of this variable with that of absolute monitoring accuracy. Measures of relative and absolute accuracy have been shown to be independent of each other (Schraw, 2009). Relative accuracy is calculated at the student level, absolute accuracy at the text level.

**Table 2** Scoring of Absolute Comprehension-Based Accuracy and Absolute Monitoring-Regulation Consistency

|   | Teacher's judgment of student's test score (monitoring-regulation consistency)/(Number of correct relations in the test (comprehension-based)) | Regulation accuracy score when teacher's choice was: 'yes, reread this text' | Regulation accuracy score when teacher's choice was: 'no, do not reread this text' |
|---|--|--|--|
| 0 | 1  | 0  | 0  |
| 1 | 0.75   | .25  | .25  |
| 2 | .50  | .50  | .50  |
| 3 | .25  | .75  | .75  |
| 4 | 0  | 1  | 1  |

This variable was recoded in our analyses (closer to 0 is more accurate) to make the interpretation of absolute monitoring accuracy and absolute regulation accuracy the same

## Quantitative analysis

Multilevel (mediation) models were run using Mplus version 7.0 (Muthén & Muthén, 1998–2015).  $K^2$  is reported as effect size indicator for the significant indirect effects (cf. Preacher & Kelley, 2011).  $K^2$  “is interpreted as the proportion of the maximum possible indirect effect that could have occurred, had the constituent effects been as large as the design and data permitted.” (Preacher & Kelly, 2011, p.106). As suggested by Preacher and Kelly (2011), we use the guidelines of Cohen (1988) to interpret the  $K^2$  values (0.01 is small, 0.09 is medium, and 0.25 is large; cf. Cohen, 1988, p. 79–81). To calculate the variance decomposition, we used a two-level model (students within school classes/teachers<sup>5</sup>). In all other analyses, we used the TYPE = COMPLEX option to take the nested structure of the data into account. We used maximum likelihood estimation and we report bootstrap confidence intervals for the mediation analysis. For all analyses, we used a Bonferroni-corrected  $\alpha$  of 0.025 (we did two comparisons with the same data: absolute and relative accuracy).

## Qualitative analysis

To investigate what performance cues, student cues, or task cues teachers used when judging students’ comprehension, qualitative analysis of the 125 think-aloud protocols was conducted ( $n_{\text{name-only condition}} = 39$ ;  $n_{\text{name-diagram condition}} = 45$ ;  $n_{\text{diagram-only condition}} = 40$ ).<sup>6</sup> The cue types performance cues (i.e., cues concerning students’ performance, derived from the diagram students made), student cues (i.e., cues concerning the student as a person, derived from teachers’ knowledge about the specific student), and task cues (i.e., cues concerning the experimental materials, that is, the texts or test questions) served as sensitizing concepts, by means of which a general sense of reference and guidance in approaching our data was provided (Blumer, 1956; Bowen, 2006). The transcripts were coded in three steps: open, axial, and selective coding (Corbin & Strauss, 1990; cf. Boeije, 2005) in NVivo version 11.4 (QSR International, 2016). In the first step (i.e., open coding) we coded utterances as student cues, task cues, or performance cues. Within these categories, we assigned codes based upon the content of the cues resulting in 33 subcategories. In the second step, the transcripts were coded axially (i.e., categories were generated/deleted/merged). We found that multiple subcategories (e.g., students’ correct and incorrect diagram answers) could be grouped under a one subcategory (e.g., correct boxes), and that one subcategory (e.g., task approach) consisted of further subcategories (e.g., students’ *general* task approach in class and students’ task approach during the *experiment*). This resulted in three main categories of cues (student-, task-, and performance cues), 27 subcategories, and six sub-subcategories.

Finally, selective coding was employed by reviewing the transcripts a final time and selecting data that exemplified the categories. We identified three cues that conceptually could not be categorized as a performance cue, student cue, or task cue. These were ‘comparing students’ performance with another student’s performance’, ‘fabricated student cue’ (i.e., cue concerning the student as a person, but that was fabricated by the teacher

<sup>5</sup> Given that each teacher participated with one class only, the class level is at the same time the teacher level.

<sup>6</sup> Accidentally, three teachers made one judgment in the name-only condition and three judgments in the name + diagram condition while thinking out-loud; moreover, one teacher did not make judgments in the diagram-only condition while thinking out-loud.

as s/he was not familiar with the student), and 'teacher' (i.e., cue concerning teachers' feelings related to the student or students' performance), which we made subcategories of a fourth category 'Miscellaneous'. Our final coding scheme thus consisted of four categories, 27 subcategories, and three sub-subcategories (see Table 3). To establish interrater reliability, 181 utterances (15.2% of all coded utterances) were randomly selected and coded independently by two coders (cf. Leighton, 2017). High inter-rater reliability was found (Krippendorff's  $\alpha=0.82$ ). In case of disagreement, raters reached consensus on the coding through discussion.

### Additional explorative analyses

When analyzing the think aloud protocols regarding teachers' cue-utilization, we noticed that teachers also often gave estimations of the cue manifestations. That is, they did, for example, not only indicate that they looked at the number of correct relations in a student's diagram, but they also indicated how many relations they thought were correct. To further explain differences between conditions in monitoring accuracy, we checked the teachers' estimations of the cues that were expected to be most diagnostic, that is, the number of correct relations, commission errors and completed boxes in the diagram.<sup>7</sup> For this purpose, we first coded for all utterances that were previously coded as referring to one of these cues, whether or not the accuracy of the teachers' estimation could be checked. Judgments that could be checked were concrete judgments for which it was clear to what information in the diagram it referred (e.g., "It was very dangerous'. This is not correct"). Judgments that could not be checked were unspecific judgments (e.g., "I think he did quite good"). To determine the interrater reliability, 60 utterances were coded by the first and second author. This resulted in a Krippendorff's  $\alpha$  of 0.79, which is considered good (Krippendorff, 2004). For completeness of the diagram, 91 of the 181 utterances could be coded regarding accuracy; for commission errors 14 of the 23; and for correct boxes in the diagram 89 of the 238. Then, for those utterances for which the accuracy could be checked, we determined the accuracy of the teachers' estimations of the cues by comparing the teachers' estimations to coded diagrams. Sixty diagrams were coded independently by two coders. This resulted in a Krippendorff's  $\alpha$  of 0.68, which is considered sufficient by Krippendorff (2004).

## Results

In Table 4, descriptive statistics for students' test scores, teachers' judgments and restudy decisions, and teachers' judgment and regulation accuracy per condition are provided. Before testing our hypotheses, we explored the variance decomposition of teachers' monitoring and regulation accuracy over the different levels. For all relative accuracy measures, the majority of the variance was located at the lowest level (the student level: 73.91%–100%). For all absolute accuracy measures, the majority of the variance was located at the lowest level (the text level; 78.94%—94.06%). This indicates that monitoring and regulation accuracy differed highly from student to student (relative) and from judgment to judgment (absolute). Differences between conditions were small, except that there was relatively more variance at the student level in the diagram-only condition for the absolute measures (see [Supplemental material](#)).

<sup>7</sup> We did expect the number of question marks to be also diagnostic but we did not check teachers' estimations of this cue given that the actual number of question marks in a student's diagram is rather obvious.



Table 3 Cue Coding Scheme

| Performance cues              | Cues concerning the performance of students, derived from the diagram students completed   | Example  |
|-------------------------------|--|--|
| Commission error              | Content that was not mentioned in the texts  | "It looks like the student made some things up himself."                               |
| Comparing with other texts    | Students' performance compared with his/her performance on another text  | "Music makes smart; she performed somewhat less than on the previous text."            |
| Completed boxes               | The completeness of (answers in) students' diagram (e.g., number of compartments filled out)   | "I miss information about the connection between the two seas."                        |
| Correct boxes                 | The correctness of (answers in) students' diagram  | " <i>More plankton</i> ; yes that is correct."   |
| Formulation                   | The formulation of (answers in) students' diagram (e.g., vague, specific)  | "The formulation might be logic, but she described it somewhat complex."               |
| Handwriting                   | Students' handwriting (e.g., neat, sloppy)   | "It's hard to read his handwriting."   |
| Notation                      | The notation of (answer in) students' diagram (e.g., arrows, spelling, strike-through)   | "I see she struck through some answers and used arrows to rearrange the compartments." |
| Question marks                | The question marks the student wrote down  | "Wow, he notated four questions marks here."   |
| Use compartments              | Students' use of the diagrams' compartments (e.g., repetition)   | "I think what that could have been placed in one compartment."                         |
| Used time                     | The time used for completing the diagram   | "I see she used 15 min to complete the diagrams."                                      |
| <b>Student cues</b>           | <b>Cues concerning the student as a person, derived from teachers' knowledge about the specific student</b>  | <b>Example</b>   |
| Approach                      | The students' task approach ...  | "She always tries her best."   |
| General                       | ... in general (e.g., in the teachers' course)   | "He likes showing off in these things; I expect he worked seriously."                  |
| Experiment                    | ... in the experiment  | "In general, he is not a fast worker."   |
| Working pace                  | ... with regard to his/her working pace (e.g., slow, fast)   | "This student has dyslexia."   |
| Behavioral/learning condition | Students' behavioural/learning condition (e.g., dyslexia, performance anxiety)   | "He is turbulent in my lessons, and can hardly concentrate."                           |
| Concentration                 | Students' concentration  | "Lisa performs mediocre in my history lessons."  |
| General skills and knowledge  | Students' general content-related skills and knowledge (e.g., remembering facts, skills related to teachers' course, students' performance on other tasks about one of the text topics, students' general reading comprehension ability) |  |
| Interest                      | Students' interest (e.g., whether it matches the texts' content)   | "Botox is an appealing text for her."  |

**Table 3** (continued)

|                              |   |   |
|------------------------------|---|---|
| Mental capacity              | Students' mental capacity (e.g., intelligence)  | "Sara is a smart girl."   |
| Motivation                   | Students' motivation  | "Well, Jill is motivated in general."   |
| Personal traits              | Students' personal traits (e.g., gender))   | "He is really something: headstrong and stubborn."  |
| <b>Task cues</b>             | <b>Cues concerning the materials of the experiment</b>  | <b>Example</b>  |
| Task type                    | Task type (relations/facts) and their connection  | "I think it is hard to answer questions about relations than about facts; so I think he will answer more fact-questions correct." |
| Test characteristics         | Teachers' ideas about the test; e.g., problem definition  | "It depends on the problem definition."   |
| Text characteristics         | Characteristics of the texts (e.g., difficulty level)   | "Money does not bring happiness was the hardest text."  |
| Text order                   | Text's order (e.g., students read the text first or last)   | "I think he knows more of that text compared to later texts."   |
| General                      | Premises about the task   | "I think it is impossible that a student answers everything correctly"  |
| <b>Miscellaneous</b>         | <b>Cues that could not be categorized as performance cue, student cue, or task cue</b>                      | <b>Example</b>  |
| Fabricated student cue       | The student as a person, but that fabricated by the teacher   | "I think this might be a girl, so she will perform well on this text."  |
| Teacher                      | Teachers' feelings related to the student or students' performance (e.g., familiarity, motivation to judge) | "I get so frustrated when a student does not even try to write down an answer."   |
| Comparing with other student | Students' performance compared to <i>another student's</i> performance                                      | "When I think about the diagrams of Carly I just saw, I expect Darryl to perform good as well"                                    |

**Table 4** Descriptive Statistics for Students' Test Scores, Teachers' Monitoring Judgments and Regulation Decisions, and Monitoring and Regulation Accuracy of Students' Comprehension

|   | Only student cues<br>(name-only<br>condition) |      | Student and<br>performance cues<br>(name + diagram<br>condition) |      | Only perfor-<br>mance cues<br>(diagram-only<br>condition) |      |
|---|---|------|--|------|---|------|
|   | M   | SD   | M  | SD   | M   | SD   |
| Students' test scores (0–4)                                   | 1.45  | 1.17 | 1.37   | 1.12 | 1.13  | 1.22 |
| Teachers judgments of students' test scores (0–4)             | 2.27  | .93  | 2.07   | 1.13 | 2.16  | 1.23 |
| Teachers' restudy decisions (0=no, 1=yes)                     | .24   | .43  | .30  | .46  | .29   | .45  |
| Absolute monitoring accuracy <sup>a</sup>                     | 1.24  | .93  | 1.13   | .91  | 1.40  | 1.19 |
| Absolute monitoring-regulation consistency <sup>b</sup>       | .43   | .23  | .38  | .26  | .35   | .27  |
| Absolute comprehension-based regulation accuracy <sup>b</sup> | .57   | .32  | .53  | .32  | .55   | .37  |
| Relative monitoring accuracy <sup>c</sup>                     | .26   | .62  | .30  | .65  | .06 <sup>e</sup>  | .67  |
| Relative monitoring-regulation consistency <sup>d</sup>       | -.61  | .69  | -.68   | .63  | -.79  | .53  |
| Relative comprehension-based regulation accuracy <sup>d</sup> | -.10 <sup>f</sup>                             | .73  | -.31   | .76  | -.29  | .73  |

<sup>a</sup>Closer to 0 is more accurate. Possible range 0 – 4

<sup>b</sup>Closer to 0 is more accurate. Possible range 0 – 1

<sup>c</sup>Closer to +1 is more accurate

<sup>d</sup>Closer to -1 is more accurate

<sup>e</sup>Does not differ significantly from zero  $t(485)=1.83, p=.07$

<sup>f</sup>Does differ significantly from zero  $t(539)=-3.13, p=.002$

### Teachers' monitoring accuracy (RQ1)

In Table 1, the findings are summarized per hypothesis. Contrary to what we expected (H1.1), we did not find significant differences in relative monitoring accuracy ( $b=-0.20, SE=0.10, p=0.04^8$ ) or for absolute monitoring accuracy ( $b=0.16, SE=0.09, p=0.053$ ) between the name-only and the diagram-only condition. Moreover, opposite to what we expected (H1.2), teachers' relative ( $b=0.24, SE=0.11, p=0.02$ ) and absolute ( $b=-0.27, SE=0.08, p=0.001$ ) monitoring was more accurate in the name + diagram condition than in the diagram-only condition. As for the explorative question regarding differences in teachers' monitoring accuracy between the name-only condition and name + diagram condition (E1.3), there were no significant differences between those conditions in either relative ( $b=0.04, SE=0.09, p=0.67$ ) or absolute monitoring accuracy ( $b=-0.11, SE=0.09, p=0.21$ ).<sup>9</sup>

### Teachers' cue-utilization (RQ2)

Table 5 shows teachers' cue-utilization per condition. Teachers used, on average, 4.18 different cues per student in the name-only condition ( $SD=2.06, \text{Min}=0, \text{Max}=10$ ), 5.87

<sup>8</sup> We used a Bonferroni corrected  $\alpha$  of .025.

<sup>9</sup> Including the number of months the teacher had known their class ( $M=7.21, SD=3.75$ ) and the class size ( $M=27.49, SD=4.93$ ) in the analyses pertaining to the name-only and name + diagram condition as covariates on teachers' monitoring accuracy did not change the conclusions.

different cues in the name + diagram condition ( $SD=2.16$ ,  $Min=2$ ,  $Max=11$ ), and 4.35 different cues in the diagram-only condition ( $SD=1.73$ ,  $Min=1$ ,  $Max=9$ ).

In the name-only condition, teachers mainly used, as expected, student cues (59%) and task cues (37%) when monitoring students' text comprehension. For more than half of their students, they mentioned the cues 'general skills and knowledge', 'interest' and 'text characteristics' when making judgments. In the name + diagram condition, teachers predominantly used performance cues (53%), then student cues (29%), and, to a lesser extent, task cues (15%). In the diagram-only condition, teachers mainly used performance cues (78%). Task cues (11%) and other (miscellaneous) cues (11%) were used to a similar extent in this condition. Furthermore, as expected, teachers in the diagram-only condition used fabricated student cues for 50% of the students. It seemed that teachers were actively searching for student cues when making their judgments (e.g., "Considering the fact that this student filled out the diagram completely correct, I think he or she is highly intelligent").

When having student and performance cues (name + diagram condition) or only performance cues available (diagram-only condition), teachers used the cues 'completed boxes' and 'correctly completed boxes' most frequently and to a similar extent in both conditions. The cue 'completeness diagram' was, for example, used in 73.33% (98) of the judgments in the name + diagram and 80% (85) of the judgments in the diagram-only condition.

### **Additional explorative analyses regarding teachers' monitoring accuracy**

When analyzing the think aloud protocols regarding teachers' cue-utilization, we noticed that teachers also often gave estimations of the cue manifestations. That is, they did, for example, not only indicate that they looked at the number of correct relations in a student's diagram, but they also indicated how many relations they thought were correct. To further explain differences between conditions in monitoring accuracy, we analyzed the teachers' estimations of the cues that were expected to be most diagnostic, that is, the number of correct relations and commission errors in the diagram and completed boxes in the diagram.

Teachers' judgments of the completeness of the diagrams were quite accurate both in the name + diagram condition (in 41 of the 42 cases) and in the diagram-only condition (44 of the 49 cases). Teachers were, in both conditions, also quite accurate in judging students' commission errors (name + diagram condition: in 2 of the 3 cases; diagram-only condition: in 7 of the 11 cases). However, teachers' judgments of the correctly completed boxes in students' diagrams was only accurate in 21 of the 45 cases (47.73%) in the name + diagram condition and only in 17 of the 44 cases (38.64%) in the diagram-only condition. For example, one teacher argued about the correctly completed boxes of a students' diagram concerning the text "Music makes smart". In this example, the student wrote in her diagram: *You can concentrate better*. When making the judgment, the teacher stated: "Yes, that is correct.". Although the teacher stated that the student provided a correct answer, it was in fact not correct because concentration was not put forward in the text, thus making the teachers' estimation of the cue inaccurate.

### **(In)direct effects of cue-availability on teachers' regulation accuracy (RQ3)**

The descriptive statistics per condition for monitoring-regulation consistency are presented in Table 4. Relative monitoring-regulation consistency was relatively high in all conditions. That is, there was a strong relation between teachers' monitoring decisions and restudy decisions (ranging from -0.61 to -0.79; values of -1 indicate perfect accuracy).

Table 5 Teachers' Cue-Utilization per Condition

|                               | Name-only condition<br>39 students          |                                      | Name+diagram condition<br>45 students       |                                      | Diagram-only condition<br>40 students       |                                      |
|-------------------------------|---|--------------------------------------|---|--------------------------------------|---|--------------------------------------|
|                               | No (%) of students the cue is mentioned for | Total no of times a cue is mentioned | No (%) of students the cue is mentioned for | Total no of times a cue is mentioned | No (%) of students the cue is mentioned for | Total no of times a cue is mentioned |
| Performance cues              |   |                                      |   |                                      |   |                                      |
| Commission error              | -   | -                                    | 5 (11.11)                                   | 5                                    | 7 (17.50)                                   | 12                                   |
| Comparing with other texts    | -   | -                                    | 8 (17.78)                                   | 8                                    | 4 (10)                                      | 6                                    |
| Completed boxes               | -   | -                                    | 33 (73.33)                                  | 98                                   | 32 (80)                                     | 85                                   |
| Correct boxes                 | -   | -                                    | 37 (82.22)                                  | 123                                  | 39 (97.50)                                  | 125                                  |
| Formulation                   | -   | -                                    | 19 (42.22)                                  | 30                                   | 16 (40)                                     | 28                                   |
| Handwriting                   | -   | -                                    | 7 (15.56)                                   | 12                                   | 6 (15)                                      | 10                                   |
| Notation                      | -   | -                                    | 10 (22.22)                                  | 13                                   | 8 (20)                                      | 13                                   |
| Question marks                | -   | -                                    | 12 (26.67)                                  | 26                                   | 3 (7.50)                                    | 3                                    |
| Use compartments diagram      | -   | -                                    | 8 (17.78)                                   | 8                                    | 9 (22.50)                                   | 16                                   |
| Used time                     | -   | -                                    | 4 (8.89)                                    | 4                                    | 1 (2.50)                                    | 2                                    |
| Sub total                     |   |                                      |   | 208 (53.47%)                         |   | 300 (77.92%)                         |
| Student cues                  |   |                                      |   |                                      |   |                                      |
| Approach                      |   |                                      |   |                                      |   |                                      |
| general                       | 10 (25.64)                                  | 12                                   | 12 (26.67)                                  | 16                                   | -   | 28                                   |
| task                          | 8 (20.51)                                   | 9                                    | 11 (28.44)                                  | 25                                   | -   | 33                                   |
| working Pace                  | 4 (10.26)                                   | 4                                    | 3 (6.67)                                    | 3                                    | -   | 7                                    |
| Behavioral/learning condition | 5 (12.82)                                   | 6                                    | 1 (2.22)                                    | 1                                    | -   | 7                                    |
| Concentration                 | 10 (25.64)                                  | 12                                   | 4 (8.89)                                    | 4                                    | -   | 16                                   |
| General skills and knowledge  | 26 (66.67)                                  | 43                                   | 16 (35.56)                                  | 20                                   | -   | 63                                   |
| Interest                      | 25 (64.10)                                  | 51                                   | 11 (24.44)                                  | 19                                   | -   | 70                                   |
| Mental capacity               | 6 (15.38)                                   | 6                                    | 10 (22.22)                                  | 10                                   | -   | 16                                   |
| Motivation                    | 1 (2.56)                                    | 1                                    | 2 (4.44)                                    | 4                                    | -   | 5                                    |
| Personal traits               | 14 (35.90)                                  | 17                                   | 8 (17.78)                                   | 11                                   | -   | 28                                   |
| Sub total                     |   | 161 (58.55%)                         |   | 113 (29.05%)                         |   | 273                                  |

**Table 5** (continued)

| Task cues                    | Name-only condition<br>39 students          |                                      | Name + diagram condition<br>45 students     |                                      | Diagram-only condition<br>40 students       |                                      | Total |
|------------------------------|---|--------------------------------------|---|--------------------------------------|---|--------------------------------------|-------|
|                              | No (%) of students the cue is mentioned for | Total no of times a cue is mentioned | No (%) of students the cue is mentioned for | Total no of times a cue is mentioned | No (%) of students the cue is mentioned for | Total no of times a cue is mentioned |       |
| Task type                    | 3 (7.69)                                    | 3                                    | 3 (6.67)                                    | 4                                    | 4 (10)                                      | 4                                    | 11    |
| Test characteristics         | 2 (5.13)                                    | 2                                    | 6 (13.33)                                   | 9                                    | 5 (12.50)                                   | 8                                    | 19    |
| Text characteristics         | 29 (74.36)                                  | 83                                   | 19 (42.22)                                  | 28                                   | 13 (32.50)                                  | 19                                   | 130   |
| Text order                   | 10 (25.64)                                  | 13                                   | 7 (15.56)                                   | 14                                   | 5 (12.50)                                   | 11                                   | 38    |
| General                      | 2 (5.13)                                    | 2                                    | 2 (4.44)                                    | 3                                    | 0   | 0                                    | 5     |
| Sub total                    |   | 103 (37.45%)                         |   | 58 (14.91%)                          |   | 42 (10.90%)                          | 203   |
| Miscellaneous                |   |                                      |   |                                      |   |                                      |       |
| Fabricated student cue       | -   | -                                    | -   | -                                    | 20 (50)                                     | 40                                   | 40    |
| Teacher                      | 2 (5.13)                                    | 2                                    | 3 (6.67)                                    | 5                                    | 2 (5)                                       | 2                                    | 9     |
| Comparing with other student | 6 (15.38)                                   | 9                                    | 5 (11.11)                                   | 5                                    | 1 (2.5)                                     | 1                                    | 15    |
| Sub total                    |   | 11 (4%)                              |   | 10 (2.57%)                           |   | 43 (11.17%)                          | 64    |
| Total                        |   | 275                                  |   | 389                                  |   | 385                                  | 1167  |

The percentages indicated for the sub totals refer to the percentage of cue type used compared to the total number of cues mentioned

Teachers' absolute monitoring-regulation consistency was moderate. That is, the difference between teachers' absolute monitoring judgments and their restudy decisions was moderate (between 0.35 and 0.43; values of 0 indicate perfect accuracy).

As expected (H3.1), higher relative monitoring accuracy predicted higher relative comprehension-based regulation accuracy ( $b = -0.51$ , 95% CI [-0.64 – -0.37],  $p < 0.001$ ; Fig. 4A-C). Similarly, higher absolute monitoring accuracy predicted higher absolute comprehension-based regulation accuracy ( $b = 0.13$ , 95% CI [0.11 – 0.16],  $p < 0.001$ ; Fig. 5A-C). Below, we present the comparisons between conditions regarding the effects of cue-availability on comprehension-based regulation accuracy.

### Name-only vs. diagram-only condition

Contrary to what we expected (H3.2a), there was no indirect effect of condition on relative comprehension-based regulation accuracy via relative monitoring accuracy ( $b = -0.11$ , 95%CI [-0.23 – -0.03],  $p = 0.04$ ; Fig. 4A).<sup>10</sup> There was, as expected (H3.3a) a significant direct effect of condition on relative comprehension-based regulation accuracy ( $b = 0.31$ , 95%CI [0.17 – 0.46],  $p < 0.001$ ; Fig. 4A), indicating that when teachers only had a student's diagram available, their relative comprehension-based regulation was more accurate than when having only a student's name available.

For absolute accuracy, there was – contrary to what we expected (H3.2a) – no indirect effect of condition on teachers' absolute comprehension-based regulation accuracy via their absolute monitoring accuracy ( $b = 0.02$ , 95%CI [0.00–0.04],  $p = 0.03$ ; Fig. 5A). In addition, there was no direct effect of condition on absolute comprehension-based regulation accuracy ( $b = -0.04$ , 95%CI [-0.10–0.02],  $p = 0.16$ ; see Fig. 5A; E3.3a).

### Name + diagram vs diagram-only condition

In contrast with what we expected (H3.2b), there was no indirect effect of condition on relative comprehension-based regulation accuracy ( $b = -0.12$ , 95% CI [-0.23 – -0.01],  $p = 0.03$ <sup>11</sup>; Fig. 4B). In addition, there was no direct effect of condition on relative comprehension-based regulation accuracy ( $b = 0.11$ , 95% CI [0.17 – 0.46];  $p = 0.28$ , Fig. 4B; E3.3b).

For absolute accuracy, there was a significant small to medium indirect effect of condition on comprehension-based regulation accuracy via absolute monitoring accuracy ( $b = -0.04$ , 95% CI [-0.09 – -0.02],  $p = 0.001$ ,  $K^2 = 0.05$ ; Fig. 5B). Contrary to what we expected (H3.2b), teachers' absolute monitoring was more accurate in the name + diagram condition than in the diagram-only condition, which in turn predicted higher levels of absolute comprehension-based regulation accuracy. There was no direct effect of condition on absolute comprehension-based regulation accuracy ( $b = 0.01$ , 95% CI [-0.04 – 0.07];  $p = 0.63$ ; Fig. 5B; E3.3b).

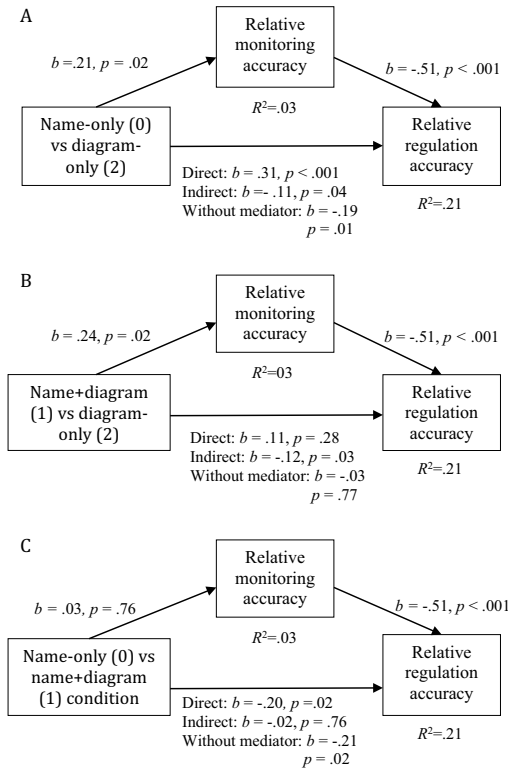
### Name-only vs. name + diagram condition

There was no indirect effect of condition on relative comprehension-based regulation accuracy via relative monitoring accuracy ( $b = 0.02$ , 95%CI [-0.11 – 0.08],  $p = 0.76$ ; Fig. 4C; E3.2c). In addition, there was – as expected (H3.3c) – a direct effect of condition

<sup>10</sup> Because we used a Bonferroni corrected  $\alpha$  of .025.

<sup>11</sup> We used a Bonferroni corrected  $\alpha$  of .025.





The negative coefficient of the effect of monitoring accuracy on regulation accuracy indicates a positive effect, given that lower values of regulation accuracy indicate higher accuracy. Furthermore, we used a Bonferroni corrected  $\alpha$  of .025

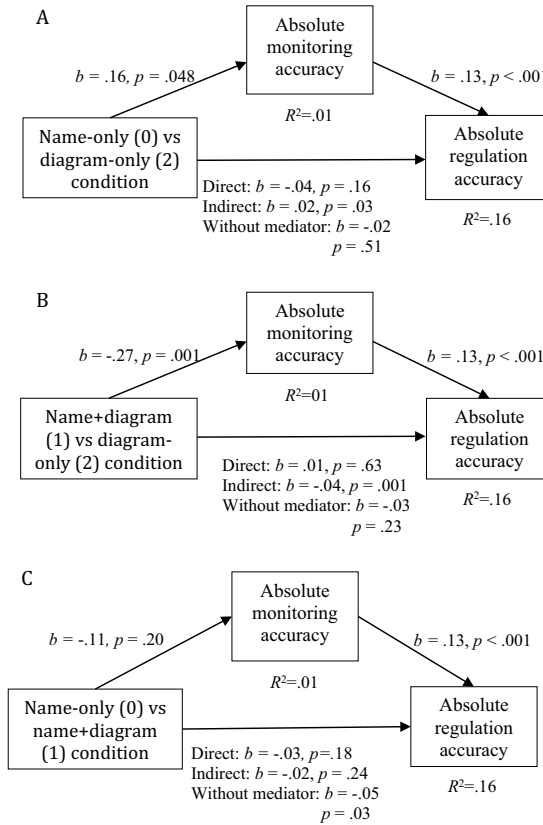
**Fig. 4** Model of Condition as a Predictor (Name-Only vs Diagram-Only [A], Name + Diagram vs Diagram-Only [B], and Name + Diagram vs Name-Only [C]) of Teachers' Relative Comprehension-Based Regulation Accuracy, Mediated by Teachers' Relative Monitoring Accuracy (Unstandardized Coefficients)

on relative comprehension-based regulation accuracy ( $b=-0.20$ , 95%CI [-0.35 – -0.01],  $p=0.02$ ; Fig. 4C). That is, teachers' relative comprehension-based regulation accuracy was higher when they had both a student's name and their diagram available compared to only a student's name.

For absolute accuracy, there was no indirect effect of condition on absolute comprehension-based regulation accuracy via absolute monitoring accuracy ( $b=-0.02$ , 95%CI [-0.04 – 0.01],  $p=0.24$ ; Fig. 5C; E3.2c). In addition, there was – contrary to what we expected (H3.3c) – no direct effect of condition on absolute comprehension-based regulation accuracy ( $b=-0.03$ , 95%CI [-0.07– 0.01],  $p=0.13$ ; Fig. 5C).

## Discussion

In the current study, we investigated: (1) how teachers' monitoring accuracy of students' text comprehension varied as a function of the availability of particular information or 'cues' (i.e., student cues only, performance and student cues, or only performance cues)



Lower values of absolute monitoring accuracy and of absolute regulation accuracy indicate higher accuracy (i.e., lower deviation of the teacher's judgment/regulation decision from the student's actual test score). Furthermore, we used a Bonferroni corrected  $\alpha$  of .025

**Fig. 5** Model of Condition as a Predictor (Name-Only vs Diagram-Only [A], Name + Diagram vs Diagram-Only [B], and Name-Only vs Name + Diagram [C]) of Teachers' Absolute Comprehension-Based Regulation Accuracy, Mediated by Teachers' Absolute Monitoring Accuracy (Unstandardized Coefficients)

(RQ1), (2) what performance, student, and task cues teachers used when monitoring students' text comprehension (RQ2), and (3) the extent to which the availability of cues directly and indirectly affected teachers' comprehension-based regulation accuracy (RQ3).

### Cue-availability, cue-utilization, and teachers' monitoring accuracy (RQ1 + RQ2)

First, we focused on teachers' monitoring, which is generally seen as the first and necessary (but not sufficient) step in the regulation process (e.g., Thiede et al., 2019). We expected that when students' work was blinded (i.e., only performance cues), teachers' monitoring of students' comprehension would be most accurate compared to having either only student cues or student cues and performance cues available. Yet, surprisingly, teachers' relative and absolute monitoring accuracy was lower when only having performance cues available (i.e., blinded student work) compared to having performance and student cues available

(i.e., students' names and their work). A possible explanation is that teachers' monitoring accuracy was *hindered* when having only performance cues available. Think aloud data showed that they might possibly have been hindered in two ways. First, as teachers fabricated student cues when only having performance cues available, the use of these fabricated student cues may have impaired their judgment accuracy. Blinding students' work thus may have had a negative effect on teachers' monitoring accuracy, perhaps not because of the availability of performance cues, but because using fabricated student cues distorted their judgments. This finding stands in contrast to the findings of Oudman et al. (2018), where teachers' monitoring was more accurate when they only had performance cues available (compared to only student cues or student and performance cues). The level of fabrication of student cues was similar in both studies (about 40% in Oudman et al., 2018, about 50% in the current study), so this cannot explain the divergence in findings between the two studies. A possible explanation for why fabrication of student cues hindered monitoring in the present study, but not in their study, might lie in the different learning tasks used: it was arguably more straightforward for teachers in the Oudman et al. (2018) study to interpret the students' performance on practice tasks, given that these were decimal magnitude problems on which students' answers were clearly right or wrong. So there would have been less room for the fabricated student cues to cloud their judgment.

Second, and related to this ease of interpretation issue, teachers regularly seemed to have difficulties with accurately interpreting the students' diagrams. Especially the number of correct relations in students' diagrams, which can be expected to be highly diagnostic for their test performance, was misinterpreted in the majority of the teachers' judgments. And this was somewhat more so when teachers only had performance cues at their disposal than when they additionally had student cues available. The higher percentage of deviation when having only performance cues available may potentially explain why teachers' monitoring accuracy in this condition was lower than when having both performance and student cues available. The accuracy of cue judgments may play an important role in the monitoring process. Surprisingly, the accuracy of cue judgments has hardly been investigated (see for an exception Schnitzler et al. (2020) who mention that teachers had troubles assessing the quality of students' contributions in classroom conversations, which can also be seen as a cue) and is not addressed in models of monitoring and regulation typically used in educational research (e.g., Koriat's cue-utilization model; Koriat, 1997). However, this issue is addressed in Funder's Realistic Accuracy Model of personality judgment (Funder, 2012). In the context of judgment accuracy of other person's personality, Funder states that, amongst others, "trait-relevant, available, and detected information must be utilized correctly. A truly friendly smile must be interpreted as friendly and not misinterpreted as insincere, sarcastic, or manipulative." (Funder, 2012, p. 178). Thus, the accuracy of judging cues (in this example, the smile) plays a role in Funder's model. Based on the current findings and theories from other, related, fields, this issue seems to deserve further investigation in the context of teachers' monitoring of students' text comprehension (see also for a discussion of this issue Van de Pol et al., 2020). Because the measurement of the accuracy of teachers' judgments of the cues was somewhat limited in the current study (i.e., only the accuracy of a few cues could be checked), this should be expanded (covering more cues) in future research. To be able to measure teachers' judgment accuracy of cues, teachers should be asked to judge the manifestations of all cues that they used. Comparison of these judgments to actual cue manifestations (e.g., the actual number of correct relations in students' diagrams) would provide information on how accurately the manifestation of the cue was judged. Note though, that this would also require researchers to map the actual manifestations of the cues under study (e.g., by administering student questionnaires to

measure their actual level of motivation or interest) in future research. The cues identified in the current study provide a starting point for identifying which cues need to be covered. Furthermore, future research could include teachers' confidence in their monitoring judgments; this may help distinguish between judgments made from cue-utilization (i.e., when teachers are highly confident) and judgments made because they have to choose (i.e., when teachers are very unconfident). In addition, a scale that asks teachers to indicate to which extent each cue has been used, could provide further information on the weighting of several cues. Finally, measuring to what extent teachers are used to this type of tasks and/or providing a more extensive training in teaching this type of tasks may prove useful in future research to be able to control for task experience or to promote teachers' familiarity with the task.

A surprising finding was that teachers' relative and absolute monitoring judgments were equally accurate when having only student cues available compared to having both student and performance cues available or only performance cues. Possibly, teachers' difficulties with accurately interpreting the performance cues hampered their monitoring accuracy when having both student and performance cues available. Still, it is quite surprising that when only having student cues available, teachers' monitoring judgments were relatively accurate, given that student cues are generally considered non-diagnostic. These findings also underline that it would be relevant for future research to start measuring the diagnosticity of a wider variety of cues to find out whether some student cues could perhaps be somewhat diagnostic.

### **(In)direct effects of cue-availability on monitoring and comprehension-based regulation accuracy (RQ3)**

Another novel contribution of the present study is that we investigated how accurate comprehension-based regulation is brought about and to what extent accurate monitoring is indeed a necessary first step to arrive at accurate regulation (cf. Thiede et al., 2019). We found that monitoring accuracy strongly predicted teachers' comprehension-based regulation accuracy, both when looking at relative and at absolute accuracy. That is, when teachers' monitoring judgments of students' understanding were accurate, their regulation decisions also often matched students' comprehension. Furthermore, we investigated to what extent the availability of cues indirectly or directly affected teachers' comprehension-based regulation accuracy via their monitoring accuracy. Teachers' comprehension-based regulation accuracy appeared to be affected both in directly and indirectly. When having only diagnostic performance cues or having both diagnostic performance cues and (non-diagnostic) student cues available, teachers were better able to indicate which texts needed to be restudied relative to other texts than when having only (non-diagnostic) student cues available (i.e., direct effect on relative comprehension-based regulation accuracy). In addition, when having both student and performance cues available, teachers' absolute monitoring judgments were more accurate which, in turn, resulted in decisions that matched students' actual comprehension better than when having only performance cues available (i.e., indirect effect on absolute comprehension-based regulation accuracy). These results partly support the claim that accurate monitoring is a necessary (but not sufficient) condition to arrive at accurate regulation (e.g., Thiede et al., 2019). That is, we found that teachers' (absolute comprehension-based) regulation was more accurate when having both student and diagram cues available compared to only diagram cues, but this was only the case when teachers' absolute monitoring was also more accurate. Yet, the direct effect

indicates that teachers' regulation accuracy is also directly affected by cue-availability, regardless of the accuracy of their monitoring judgments. Possibly, teachers use different cues for monitoring and regulation decisions, or they give different weight to different cues. Future research, in which the regulation process is further investigated for example by asking teachers to think out loud, may give more insight in whether teachers make judgment and regulation decisions in different ways, for example by using different cues. Such future research could focus on teachers, but it would also be interesting to compare judgments made by teachers to judgments made by others (who might either be involved in regulating students' learning sometimes, like parents or peers, or who might be complete strangers). To what extent would others also benefit from having access to performance cues and what other factors play a role in accurate monitoring? Peers have experience with the task but not necessarily with the scoring criteria of the test in contrast to teachers who have experience with both. Furthermore, parents have no experience with the task and criteria but (like teachers) know their children, whereas strangers have no experience with the task, criteria, and no knowledge of the children. The results of the current study show that knowing the student one is judging, is an important factor in bringing about accurate monitoring. Further research that compares monitoring accuracy and cue-utilization of teachers to other groups, may give us more insight into factors that enhance or hamper accurate monitoring judgments. A more general point is that not all findings were the same for relative and absolute monitoring and regulation accuracy. Making absolute monitoring decisions (i.e., deciding per text how well a student understands it) benefitted from having both performance cues and student cues compared to only performance cues, which was not the case for relative monitoring decisions (i.e., deciding which text a student understands less well compared to other texts). In addition, teachers' absolute regulation was more accurate when having only performance cues compared to having only student cues, which was not the case for relative regulation accuracy. Although our data shows that teachers take these cues into account when making monitoring judgments, the role of these cues for absolute and relative measures seems to differ. From these results, it seems that knowing which student is being judged helps in arriving at more accurate absolute monitoring and regulation, but not necessarily for more accurate relative monitoring and regulation. Future research could further investigate this issue and unravel in what ways accurate relative and absolute monitoring and regulation are brought about. Including both measures in future studies can help us to better understand these processes.

## Conclusion

This study is one of the first that focused relations between availability of cues, monitoring accuracy, and regulation accuracy. Most importantly, this study suggests that accurate comprehension-based regulation may be brought about in different ways than accurate monitoring, possibly due to diverging cue use in monitoring and regulation.

Provided that the results of the current study would be replicated in other samples and different tasks and appear to be robust, models of teacher monitoring and regulation may further refine the interrelations between these two processes, acknowledging that the availability of cues may affect monitoring and regulation differently. In addition, teacher monitoring and regulation models may additionally take into account that regulation accuracy can be brought about both in direct and indirect ways. Furthermore, the results of the current study may be helpful in designing interventions for promoting teachers'

monitoring and regulation accuracy. For instance, such interventions might focus on promoting deep processing of diagnostic cues to foster both monitoring and regulation accuracy, making accurate judgments of diagnostic cues, and possibly inhibiting the fabrication of student cues when inspecting blinded student work.

## Appendix A<sup>12</sup>

One of the texts students read, titled “The Suez Canal”.

“The Suez Canal, which connects the Indian Ocean and the Mediterranean Sea with each other, is of great importance to the world. Originally, there was no natural water connection between the Atlantic and the Indian Ocean. Between these two seas is a desert. This meant that trading ships that traveled from the harbor city Jeddah in Saudi Arabia to Europe had to make a long journey around the whole African continent. It was therefore decided that a shorter waterway was needed that would connect the two oceans with each other. For this reason, the Suez Canal, which was designed by the Austrian engineer Alois Negrelli, was dug. For years, workers were digging; the canal was finally opened in 1869 for shipping. By the digging of the Suez Canal, the distance from the harbor city of Jeddah to the harbor city of Rotterdam has been reduced by 40%. Through the Suez Canal, the distance between these cities is 6,337 nautical miles, when ships sail around the African continent this distance is 10,743 nautical miles.”

## Appendix B<sup>12</sup>

Question about causal relations for the text ‘The Suez Canal’.

The distance for trading ships that sail between Jeddah and Rotterdam has been reduced a lot. For what reasons has the distance between Jeddah and Rotterdam been reduced?

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1007/s11409-021-09268-6>.

**Acknowledgements** We would like to thank all students who helped collecting and coding the data.

**Funding** During the realization of part of this research the first author was funded by a Veni grant from the Netherlands Organization for Scientific Research awarded to the first author (grant number: 451–16–012).

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

## Declarations

**Ethics approval** We complied with the APA ethical standards for treatment of human participants, informed consent, and data management. Furthermore, at the time of data gathering for the current study (2016), ethical approval was not obliged at Utrecht University, especially not for non-invasive studies containing only

<sup>12</sup> Appendices are reprinted with permission from “Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy”, by M. van Loon et al. (2014), *Acta Psychologica*, 121, p. 153.

tasks that are similar to everyday school tasks. Note though, that a later study with a comparable design and procedure, was approved by the Utrecht University ethics committee.

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

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

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

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Artelt, C., & Rausch, T. (2014). Accuracy of Teacher Judgments. In S. Krolak-Schwerdt, S. Glock & M. Böhmer (Eds.), *Teachers' professional development. Assessment, Training, and Development* (pp. 27–43). Rotterdam/Boston/Tapei: SensePublishers.
- Baars, M., van Gog, T., de Bruin, A., & Paas, F. (2014). Effects of problem solving after worked example study on primary school children's monitoring accuracy. *Applied Cognitive Psychology*, 28(3), 382–391. <https://doi.org/10.1002/acp.3008>
- Blumer, H. (1956). Sociological analysis and the "variable". *American Sociological Review*, 21, 683–690. <https://doi.org/10.2307/2088418>
- Boeije, H. (2005). *Analyseren in kwalitatief onderzoek [Analyzing in qualitative research]*. Boom Lemma Uitgevers.
- Bowen, G. A. (2006). Grounded theory and sensitizing concepts. *International Journal of Qualitative Methods*, 5, 12–23. <https://doi.org/10.1177/160940690600500304>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.
- Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as Judgment-in-Context: Analysing how teachers evaluate students' writing I. *Educational Research and Evaluation*, 13, 401–434. <https://doi.org/10.1080/13803610701728311>
- Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, 13, 3–21. <https://doi.org/10.1007/BF00988593>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.
- De Bruin, A., Van de Pol, J., Van Loon, M., & Van Gog, T. (2017). Students' and teachers' monitoring of causal relations texts: Do diagrams help? Paper session presented at the Psychonomics Society Meeting, Vancouver, US
- Dunlosky, J., 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>
- Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction*, 24, 58–61. <https://doi.org/10.1016/j.learninstruc.2012.05.002>
- Engelen, J. A., Camp, G., van de Pol, J., & de Bruin, A. B. (2018). Teachers' monitoring of students' text comprehension: can students' keywords and summaries improve teachers' judgment accuracy?. *Metacognition and Learning*, 13(3), 287–307. <https://doi.org/10.1007/s11409-018-9187-4>
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as Data*. MIT Press.
- Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science*, 21(3), 177–182. <https://doi.org/10.1177/0963721412445309>
- 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: Cambridge University Press. <https://doi.org/10.1017/9781108235631.025>
- Hardy, I., Decristan, J., & Klieme, E. (2019). Adaptive teaching in research on learning and instruction. *Journal for Educational Research Online*, *11*(2), 169–191.
- Hayes, A. F., & Coultts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. *But... Communication Methods and Measures*, *14*, 1–24. <https://doi.org/10.1080/19312458.2020.1718629>
- Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research*, *30*, 411–433. <https://doi.org/10.1111/j.1468-2958.2004.tb00738.x>
- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology*, *126*, 349–370. <https://doi.org/10.1037/0096-3445.126.4.349>
- Leighton, P. J. (2017). Think aloud interviews procedures. In: *Using think-aloud interviews and cognitive labs in educational research*. Oxford: Oxford University Press.
- Muthén, L. K., & Muthén B. O. (1998–2015). *Mplus user's guide*, 7th Ed. Los Angeles, CA: Muthén & Muthén.
- Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of Knowing predictions. *Psychological Bulletin*, *95*(1), 109–133. <https://doi.org/10.1037/0033-2909.95.1.109>
- Oudman, S., Van de Pol, J., Bakker, A., Moerbeek, M., & Van Gog, T. (2018). Effects of different cue-types on the accuracy of primary school teachers' judgments of students' mathematical understanding. *Teaching and Teacher Education*, *76*, 214–226. <https://doi.org/10.1016/j.tate.2018.02.007>
- Parsons, S. A., Vaughn, M., Scales, R. Q., Gallagher, M. A., Parsons, A. W., Davis, S. G., ... & Allen, M. (2018). Teachers' instructional adaptations: A research synthesis. *Review of Educational Research*, *88*(2), 205–242. <https://doi.org/10.3102/0034654317743198>
- Preacher, K. J. & Kelley, K. (2011). Effect sizes measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, *16*(2), 93–115. <https://doi.org/10.1037/a0022658>
- QSR International. (2016). NVivo qualitative data analysis Software; Version 11.4.
- Robertson, C. T., & Kesselheim, A. S. (Eds.). (2016). Blinding as a solution to bias: Strengthening biomedical science, forensic science, and law. Academic Press.
- Schnitzler, K., Holzberger, D., & Seidel, T. (2020). Connecting judgment process and accuracy of student teachers: Differences in observation and student engagement cues to assess student characteristics. *Frontiers in Education*, *5*, Article 602470. <https://doi.org/10.3389/educ.2020.602470>
- Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, *4*, 33–45. <https://doi.org/10.1007/s11409-008-9031-3>
- Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*, 743–762. <https://doi.org/10.1037/a0027627>
- Südkamp, A., Möller, J., & Pohlmann, B. (2008). Der Simulierte Klassenraum: Eine experimentelle Untersuchung zur diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie*, *22*(34), 261–276. <https://doi.org/10.1024/1010-0652.22.34.261>
- 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](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., Brendefur, J. L., Carney, M. B., Champion, J., Turner, L., Stewart, R., & Osguthorpe, R. D. (2018). Improving the accuracy of teachers' judgments of student learning. *Teaching and Teacher Education*, *76*, 106–115. <https://doi.org/10.1016/j.tate.2018.08.004>
- Thiede, K., 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: 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, Article 100374. <https://doi.org/10.1016/j.edurev.2020.100374>
- 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., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review*, *22*(3), 271–296. <https://doi.org/10.1007/s10648-010-9127-6>

- Van de Pol, J., Volman, M., Oort, F., & Beishuizen, J. (2014). Teacher scaffolding in small-group work: An intervention study. *Journal of the Learning Sciences*, 23(4), 600–650. <https://doi.org/10.1080/10508406.2013.805300>
- Van 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>
- Vandevelde, S., Van Keer, H., Schellings, G., & Van Hout-Wolters, B. (2015). Using thinkaloud protocol analysis to gain in-depth insights into upper primary school children's self-regulated learning. *Learning and Individual Differences*, 43, 11–30. <https://doi.org/10.1016/j.lindif.2015.08.027>
- Van de Pol, J., van Loon, M., van Gog, T., Braumann, S., & de Bruin, A. (2020). Diagramming as a generative activity to improve students' and teachers' monitoring and regulation of students' text comprehension: Current findings and future directions. *Educational Psychology Review*, 1-27. <https://doi.org/10.1007/s10648-020-09560-y>
- Webb, M. B. (2015). *Exploring the correlation between teachers' mindset and judgment accuracy to reveal the cues behind teachers' expectations (Doctoral dissertation)*. Boise State University.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.