Teaching and Teacher Education 104 (2021) 103386

Contents lists available at ScienceDirect

Teaching and Teacher Education

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

Research paper

The relationship between teachers' cue-utilization and their monitoring accuracy of students' text comprehension

Janneke van de Pol [a, *], Tamara van Gog [a], Keith Thiede [b]

[a] Department of Education, Utrecht University, the Netherlands
[b] College of Education, Boise State University, Boise, USA

HIGHLIGHTS

- Teachers' monitoring accuracy of their students' text comprehension was investigated.
- Performance cues, student cues, and task cues were available during monitoring.
- *Mere use* of diagnostic cues was not sufficient to promote teachers' monitoring accuracy.
- Using non-diagnostic student cues (e.g., students' extraversion) hampered teachers' monitoring accuracy.
- Accurately judging the values of one of the diagnostic cues increased teachers' monitoring accuracy.

ARTICLE INFO

Article history:
Received 19 May 2020
Received in revised form
7 May 2021
Accepted 8 May 2021
Available online 24 May 2021

Keywords:
Teacher monitoring accuracy
Teacher judgment accuracy
Student text comprehension
Cue-utilization
Cue-diagnosticity

ABSTRACT

We investigated to what extent teachers' *use* of diagnostic cues and the accuracy with which they *interpreted or judged the values* of those cues affected teachers' monitoring accuracy. Forty-six secondary education teachers judged the text comprehension of six students (216 students in total). *Mere use* of diagnostic cues appeared not sufficient. Rather, accurately judging the values of a diagnostic performance cue was related to higher monitoring accuracy. Using non-diagnostic student cues hampered teachers' monitoring accuracy. The key to further improve monitoring accuracy might lie in improving teachers' ability to accurately judge diagnostic cues and help them ignore non-diagnostic cues.

Every student is different and thus has different needs to learn effectively. Instructional support that is adapted to these needs promotes students' learning (Author, 2010; Parsons et al., 2018). To deliver adaptive support, teachers must know what their students know (Author, 2011; Klug et al., 2013). During or in between lessons, determine what their students know by looking at students' work. Based on this, teachers adapt their instruction or lesson plan for subsequent lessons. Yet, a meta-analysis showed that teachers' monitoring accuracy of students' performance (i.e., the relation between teachers' judgments of students' performance and students' actual performance) is far from perfect and that there is much room for improvement (Südkamp et al., 2012). In the current study, we focus on this essential skill of monitoring students' performance, which is a necessary condition for delivering adaptive instruction (Author, 2019B). According to the cue-utilization framework (Koriat, 1997), people use cues (i.e., "bits of information that might potentially be drawn upon or referred to by a teacher to inform a judgment" Snow, as cited in Cooksey et al., 2007, p. 431) when making judgments. Teachers for example can deduce cues by inspecting students' work (e.g., correctness of answers). Additionally, teachers can use information about students such as effort in class or interest in a text topic or information about the task (e.g., text difficulty or length).

Using cues that are predictive or diagnostic of the judged outcome (here: text comprehension) promotes teachers' monitoring accuracy. For example, when teachers focus on students' ability to explain a text (i.e., a diagnostic cue), teachers' judgments of students' test scores are more accurate than when focusing on whether students find a text interesting (i.e., a non-diagnostic cue). Previous studies in which teachers were provided with information

* Corresponding author.
E-mail address: j.e.vandepol@uu.nl (J. van de Pol).

https://doi.org/10.1016/j.tate.2021.103386

containing diagnostic cues, however, showed mixed results regarding teachers' monitoring accuracy. One study found no effect of access to diagnostic cues (Author, 2019) while other studies found a positive effect (Author, 2010, Author, 2010). Yet, teachers' monitoring accuracy was quite low in all studies.

Just making information available from which diagnostic cues can be deduced may be insufficient to boost monitoring accuracy. Teachers may not know what information to focus on or process information superficially (Glock et al., 2012) and therefore not actually *use* diagnostic cues. Moreover, even if teachers use diagnostic cues, they would have to accurately interpret or judge the actual values of those cues (i.e., 'used-cue value judgment accuracy') for their monitoring accuracy to improve. For instance, a student's ability to explain a text is a diagnostic cue, but if a teacher judges that a student can explain a text well, whereas this is actually not the case, their cue-judgment would be inaccurate.

The current study's aim is to investigate to what extent cue-utilization and used-cue value judgment accuracy are related to teachers' monitoring accuracy of their students' text comprehension. Although it may seem self-evident that used-cue value judgment accuracy is related to teachers' monitoring accuracy, nothing is known yet about this relation and previous studies only focused on cue-utilization and its relation with teachers' monitoring accuracy of students' comprehension. Determining the role of used-cue value judgment accuracy is theoretically important as this aspect may need to be added to theoretical and/or process models of teacher monitoring. Additionally, it is practically important as it may shift the focus of interventions for improving teachers' monitoring accuracy from cue-utilization to used-cue value judgment accuracy. In the current study, teachers had access to students' products of generative activities they engaged in. Generative activities refer to activities that involve "actively making sense of to-be-learned information by mentally reorganizing and integrating it with one's prior knowledge (Fiorella & Mayer, 2016, p. 717). Engaging in such activities generates diagnostic cues for students and teachers (Author, 2014, 2019bib_Author_2010bib_Author_2010). Such activities, such as making drawings or completing diagrams about a text, are common practice in education (cf. Fiorella & Mayer, 2016). In the current study, teachers viewed students' completed diagrams, as these concisely represent students' text comprehension.

1. Teachers' cue-utilization

Teachers use various cues when monitoring students' comprehension (e.g., Author, 2018). First, teachers use *performance* cues: information about students' prior performance (Table 1). For instance, students' achievement in other or the judged domain (e.g., on prior tasks or generative activities; Author, 2018; Dompnier et al., 2006; Helwig et al., 2001).

Second, teachers use *student* cues: information about students such as effort (Kaiser et al., 2013), nationality (Furnari et al., 2017; Holder & Kessels, 2017; Meissel et al., 2017), learning problems (Johnston et al., 2019), IQ, and interest (Cooksey et al., 2007; Webb, 2015). Regarding students' gender, results are mixed: in some studies teachers used gender (Holder & Kessels, 2017; Kaiser et al., 2015; Meissel et al., 2017), while in others they did not (Hecht & Greenfield, 2002; Helwig et al., 2001).

Finally, teachers use *task* cues: information about the task, such as text content or item/task difficulty (Author, 2018; Webb, 2015). Author (2020) showed that teachers used on average 5.87 cues per judgments. They mostly used performance cues (e.g., diagram correctness), followed by student cues (e.g., IQ). Task cues (e.g., text length) were used least often. According to the cue-utilization framework (Koriat, 1997), monitoring accuracy depends on how diagnostic the used cues are.

When studies measured cue-utilization, they mostly did so by calculating the correlation between cue values and judgments (e.g., Author, 2014; Schleinschok et al., 2017). Schleinschok et al. (2017), for example, calculated correlations between characteristics of drawings that students made about texts (e.g., idea units) and judgments about students' text comprehension to express cue-utilization. However, teachers have no access to these cue values; they also *judge* these values. So relating actual cue values to teachers' judgments of students' text comprehension may not necessarily express their cue-utilization. Especially given that teachers often overestimate their students, this correlational measure may overstate their cue-utilization. Therefore, we used teachers' self-reported cue-utilization. For this purpose, we compiled a cue-list based on think-aloud data of previous studies (Author, 2018; 2020, Table 1) and complemented with cues from the literature (Bennett et al., 1993; Cooksey et al., 2007; Dinsmore & Parkinson, 2013; Dusek & Joseph, 1983; Jenkins & Demaray, 2016; Mizala et al., 2015; Rausch et al., 2015; Weaver & Bryant, 1995).

2. Cue-diagnosticity

A cue is highly diagnostic when the relationship between actual cue values (e.g., commissions in students' work) and judged outcomes (e.g., students' text comprehension test score) is strong. Kostons and de Koning (2017), for example, showed that elements and details in students' drawings about texts were diagnostic of students' test performance. Moreover, drawings in their experimental condition, which aimed at and resulted in improved monitoring accuracy, contained more of these diagnostic cues than drawings in the control condition in which monitoring accuracy was lower. Measuring cue-diagnosticity can thus help explaining monitoring accuracy differences.

Generally, performance cues seem most diagnostic (e.g., Author, 2010B; 2019B; Griffin et al., 2009). Next to using diagnostic cues, non-diagnostic cues should be ignored. Using vignettes manipulating cue availability, Kaiser et al. (2015) showed that teachers' judgments of students' mathematics achievement were more accurate when they only had (diagnostic) performance cue values available (i.e., oral/written mathematics achievement) than when they additionally had student cue values available (e.g., students' gender, intelligence). When monitoring their *own* students' mathematics performance, teachers were also most accurate when having only diagnostic performance cues available (by providing teachers with anonymized student work) instead of only student cues or performance and student cues (Author, 2018).

Two studies that directly measured the performance cues' diagnosticity by relating actual cue values to students' test scores, showed that some performance cues are highly diagnostic whereas other performance cues are not (Author, 2014, 2019bib_Author_2010bib_Author_2010). Specifically, correct causal relations in students' diagrams was highly diagnostic (r = 0.40-0.50); commissions and factual information in students' diagrams had low diagnosticity ($r = -0.15$ to -0.25 and $r = -0.09$ respectively). This suggests that teachers' judgments of students' performance would be more accurate when using causal relations but not when using commissions.

3. Teachers' cue-utilization and used-cue value judgment accuracy

Few studies measured teachers' cue-utilization. Think-aloud analysis showed that the higher teachers' use of diagnostic performance cues and the lower their use of non-diagnostic student and task cues, the higher their monitoring accuracy of students'

Table 1
Cue-list (columns 1–3) and information about the instruments to measure students' actual cue values (columns 4–6).

Main category	Sub category	Cue + explanation	Measurement instrument actual cue values	Internal consistency (Ω)	Example item (nr of items; answer scale)
Performance cues	Completeness diagram	No. of omission errors (blank boxes/ question marks)	Coding scheme students' diagrams	NA	NA
		No. of boxes containing information **not** in the text (commission error)			
	Correctness diagram	No. of correct facts (non-essential info)			
		No. of correct elements/boxes			
		No. of correct cause-effect relations			
	Phrasing	Mean no. of words in diagram boxes	Count	NA	NA
	Used time	Time in minutes to complete diagram	Log	NA	NA
Student cues	Students' general attitude towards school work	Student effort in teacher's lessons (e.g., work hard, pay attention).	Ongoing Engagement Subdomain scale (IRRE, 1998).	.76	I pay attention in the lessons of teacher X (5; 1 (totally disagree) to 4 (totally agree))
		Student precision when working on assignments/tests (tidy/systematic)	Big Five conscientiousness scale (Goldberg, 1992)	.86	To what extent do you show the following traits in class of teacher X: precision (6; 1 (not true at all) to 7 (entirely true))
	General knowledge and skills students	General reading comprehension level[a]	Cloze test, developed for current study	.42	Professor Ian Neary of the University of Oxford _____ to explain this (20; open question)
		Student ability to reproduce facts	Reproduction test (Van Loon et al., 2014)	.34	What fish did the politician hold in the picture? (5; open question)
		Student achievement for teacher's subject (report card grade/current mean grade)[b]	Student self-reported grade	NA	NA
		Student achievement for other subject (report card grade/current mean grade)	Student self-reported grade	NA	NA
	Interest student	How interesting/absorbing/ fascinating does this student find the text topic?	Situational interest scale per text (Linnenbrink-Garcia et al., 2010)	1.00	The topic of this text is fascinating to me. (4 per text; 1 (not at all true) to 5 (very true)).
	General personal characteristics	Extraversion: How talkative/active is this student generally in class?	Big Five extraversion scale (Goldberg, 1992)	.89	To what extent do you show the following traits in class of teacher X: quietness (reverse coded) (6; 1 (not true at all) to 7 (entirely true))
		Degree of self-efficacy (certainty/ self-confidence) with regard to school work for teacher's subject.	Perceived self-efficacy scale (bib_Marsh_et_al_2006Marsh et al., 2006)	.83	I'm certain I can understand the most difficult material presented the study materials of subject X (4; 1 (almost never) to 4 (almost always)).
		The student's gender	Student self-report	NA	NA
		Learning problems student (dyslexia, adhd, add, autism, giftedness, dyscalculia, Dutch as second language)	Student self-report	NA	0: student does not have the learning problem//1: student has the learning problem
		Nationality student: Based on birth country student/mother/father	Student self-report	NA	5: student, mother and father born in the Netherlands (NL)//4: student and mother or father born in NL//3: student born in NL, mother and father not//2: student not born in NL, mother and father born in NL//1: student not born in NL, mother or father born in NL//0: student, mother and father not born in NL.
	Mental capacity	Student's IQ	Raven standard progressive matrices (Bilker et al., 2012)	.54	(9 items; per item 6 or 8 answer options)
Task cues	Text characteristics	No. of facts in the text	Count		
		Text length (no. of lines)	Count (Text music: 14, text metro: 12, text concrete: 13)		
		No. of difficult words in the text	Average all participants		
	Text position	First, second, or third text for this student	Log		

[a] Three assistants coded data of 93 students. With a Krippendorff's alpha of .98, the interrater reliability was good (Landis & Koch, 1977).
[b] The diagnosticity and used-cue value judgment accuracy is based on students' grades for Math, Science, and English.

performance was (Author, 2018). However, teachers' utilization of (non-)diagnostic cues could not explain differences in teachers' monitoring accuracy in another study (Author, 2020). This study found that teachers' monitoring accuracy of students' text comprehension was lower when having only performance cues available compared to having performance *and* student cues available. This finding was surprising as analyses of the think-aloud protocols showed that when only having performance cues available, teachers used up to 25% more (diagnostic) performance cues than when having performance and student cues available. Further analyses, however, suggested that, even though teachers used diagnostic cues, they had difficulties in accurately interpreting or judging cues that could be derived from the diagrams (e.g., correct relations). Thus, correct cue interpretation may also play a role (cf. Funder, 1999). To further explore this, we asked teachers to judge cue values of used cues. We compared these cue-value judgments to the actual cue values to compute the *used-cue value judgment accuracy*. To the best of our knowledge, this is the first study to

relate teachers' used-cue value judgment accuracy to teachers' monitoring accuracy of students' performance.

4. The current study

To better understand how cue-utilization relates to teachers' monitoring accuracy, we examined cues that teachers used to judge students' text comprehension and their judgments of the cue values. Teachers completed three conditions in a within-subjects design. First, teachers only judged students' performance (performance-only condition). Then, teachers judged students' performance and selected used cues from a list (judgment + cue-list condition). Finally, they judged students' performance, selected cues, and rated the perceived cue values.

Forty-six secondary school teachers judged their students' text comprehension while having various information sources available from which they could deduce cues: students' completed diagrams about causal relations in each text (giving access to performance cues), students' names (access to student cues), and the texts and test (access to task cues). Teachers had these information sources available in all three conditions. A special feature of this study is that we measured the actual values of all included cues (e.g., students' IQ, correct relations in students' diagrams, and text characteristics). This is firstly useful for future research that may use the cue diagnosticity. Additionally, measuring diagnosticity enables us to: (1) take the actual cue-diagnosticity for this sample into account in interpreting our results, and (2) measure how accurately teachers judge cues by relating the actual cue values to teachers' cue judgments. We address the following research questions:

RQ1: To what extent are a wide range of performance, student, and task cues (cf. Table 1) diagnostic of students' text comprehension? Based on previous research (Author, 2014, 2019bib_Author_2010bib_Author_2010), we expect that correct relations in students' diagrams are highly diagnostic ($r > 0.50$) whereas commissions and factual information in the diagram are low diagnostic ($r < 0.30$). The diagnosticity of other cues is explored. Moreover, we expect that performance cues are more diagnostic than student and task cues.

RQ2: What cue-use patterns occur when monitoring students' text comprehension? A cue-use pattern is the constellation of cues used for a judgment consisting of one or several cues. Based on Author (2020), we expect that teachers use − on average −six cues per judgment and mostly use performance cues, followed by student, and task cues. Cue-use patterns are explored.

RQ3: How accurately can teachers judge the cue value of performance, student, and task cues (i.e., used-cue value judgment accuracy)? Generally, teachers' judgment of student characteristics (e.g., self-concept and academic interest; Karing, 2009; Praetorius et al., 2013; 2017) and task characteristics (e.g., text/item difficulty; Hoffmann & Böhme, 2013; McElvany et al., 2009) are more accurate than of students' performance (Artelt & Rausch, 2014; Südkamp et al., 2012) so we expect that teachers judge student and task cues more accurately than performance cues.

RQ4: To what extent do cue-utilization and used-cue value judgment accuracy relate to teachers' monitoring accuracy of students' text comprehension? We expect that when teachers use highly diagnostic cues and judge these cues accurately, their judgments of students' text comprehension is most accurate.

5. Method

5.1. Participants and design

Forty-six secondary education teachers of subjects for which text comprehension is important (e.g., languages, history/geography) participated (64% female; 94% Dutch). The sample-size was based on a multilevel a-priori power analysis (power = .80) conducted in spa-ml (Moerbeek & Teerenstra, 2015). Teachers had known their classes for 10.64 months on average (SD = 6.39)[1] and had − on average − 12.5 years of teaching experience (SD = 7.92). They received a €50 voucher for participation.

The study had a within-subjects design, with all teachers judging three students' text comprehension under three conditions in the following order: judgment-only; judgment + cue-list; judgment + cue-list + cue judgment. For each condition, teachers judged three of their students' comprehension and made separate judgments for each text read by students. Overall, teachers made 405 judgments (135 students*3 texts) in the judgment-only condition, 405 judgments (135 students*3 texts) in the judgment + cue-list condition, and 408 judgments (136 students*3 texts) in the judgment + cue-list + cue judgment condition.[2] Although there were three conditions, only the judgment + cue-list and judgment + cue-list + cue-value-judgment condition provided teachers' self-reported cue-utilization data, which was this study's focus. The judgment-only condition was implemented to check whether explicating cue-utilization and judging cue-values was related to teachers' monitoring accuracy. There were no significant differences between conditions regarding students' test scores, teachers' judgments, and teachers' monitoring accuracy in terms of deviation or bias (all p's > 0.05; see Table 2 for *M's* and *SD's*). In the current study, we only used data of the judgment + cue-list and judgment + cue-list + cue-value-judgment condition (261 students; M_{age} = 15.15, SD = 1.37; 50.7% female; 93.8% born in the Netherlands).

Students whose text comprehension was judged were selected based on their general reading comprehension test scores (see student measures). For each condition, we selected a student with low (≈20th percentile), medium (≈50th percentile), and high (≈80th percentile) scores. Within each condition, the order in which these three students were judged was randomized. This study received approval from the ethics review board of the first author's institute.

6. Student measures

6.1. Expository texts

Students read three texts, derived from the study by Author (2019B). The topics of the texts were "Music makes smart" (167 words), "Sinking of metro cars" (158 words), and "Concrete constructions" (166 words). Each text contained five clauses conveying causal relations (see Appendix for instructions).

6.2. Student diagrams

After reading, students completed diagrams. For these diagrams, students were asked to write down the text's cause-and-effect relations (see Appendix for instructions). Please see Fig. 1 for an example. Students did not receive feedback on the quality of their diagrams. Coding of the diagrams, information about the interrater

[1] There was no effect of the number of months the teacher knew their class on their judgment accuracy of student characteristics in our data; test results can be requested from the first author.

[2] For some teachers, there were not enough students available (due to illness or because they declined participation); in the judgment-only condition, three teachers made judgments about two students, in the judgment + cue-list condition, one teacher made judgments about two students and one teacher about one student, and in the judgment + cue-list + cue-judgment condition, two teachers made judgments about two students.

Table 2
Means and standard deviations of students' test scores, teachers' judgments and teachers' monitoring accuracy.

	Condition							
	Overall		Judgment-only		Judgment + cue-list		Judgment + cue-list + cue-value-judgment	
	M	SD	M	SD	M	SD	M	SD
Student test score (0–8)	3.72	2.61	3.90	2.52	3.83	2.64	3.44	2.64
Teacher judgment (0–8)	4.77	2.40	4.90	2.32	4.87	2.28	4.55	2.57
Teacher monitoring accuracy – bias[a]	1.15	2.62	1.11	2.56	1.14	2.58	1.18	2.73
Teacher monitoring accuracy – absolute deviation[b]	2.19	1.84	2.12	1.80	2.21	1.75	2.25	1.95

[a] Range: 8 (underestimation) to +8 (overestimation); 0 is most accurate.
[b] Range: 0 (most accurate) to 8 (least accurate).

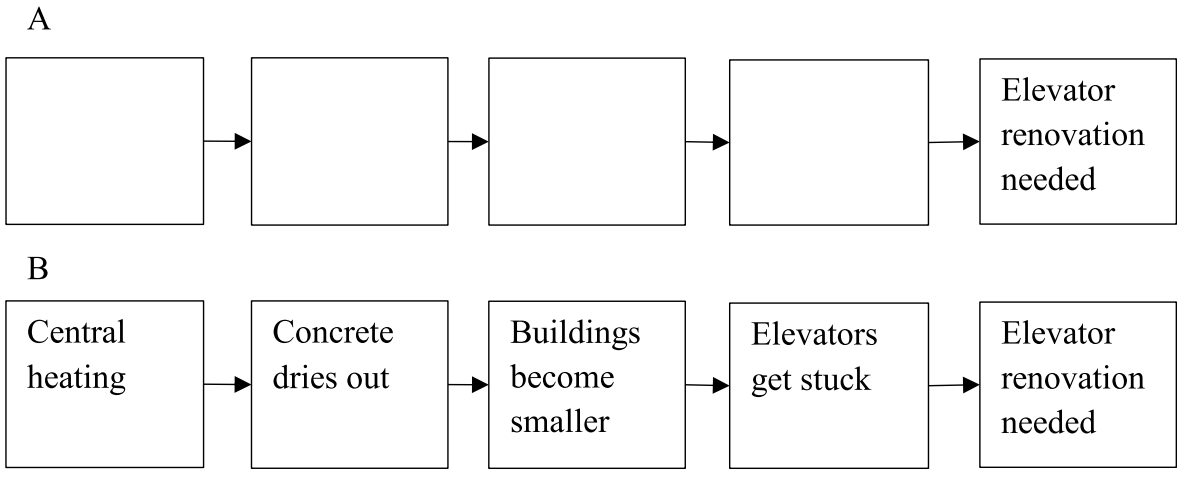

Fig. 1. An Empty and a correctly completed diagram for the text 'concrete constructions'.

reliability and an example can be found in the section 'Performance cue values – diagram cues'.

6.3. Text comprehension test

For each text, students completed a test question. Students were asked to describe (in text format) the causal relations in each text. They were provided with one of the causes or effects for each question and with signaling words that they could use to make the order of the causes and effects clear (e.g., 'for that reason', 'first'). See Appendix for instructions and an example question.

For scoring students' answers, we used an existing answer format (cf. Author, 2014; 2019). The answer format was straightforward, as it consisted of the correct cause-and-effect elements and the order of these elements as represented in the texts. Students were assigned one point for each correct element that was detected in their answer (range per text: 0–4). They did not get points for copying the provided element. Data of 50 students was double coded by two assistants and the interrater reliability was substantial (Krippendorff's alpha: 0.93). Additionally, we determined the number of correct combinations of two elements (i.e., the number of correct relations per text: 0–4) (Krippendorff's alpha: 0.88). The total test score was the sum of correct elements and relations (range: 0–8). The reliability of the test was acceptable ($\alpha = 0.73$). Furthermore, the test seemed to validly measure students' understanding of causal relations in the text. That is, just reproducing information from the text would not result in a high comprehension score; the students had to show actual understanding of the link between the causes and effects by describing them in the right order to obtain points. This is substantiated by high correlations between students' test and diagram scores indicating students' understanding of causal relations in the texts such as the correct relations ($r = 0.96$) and the correct elements ($r = 0.91$).

6.4. Actual cue values

Table 1 summarizes the most important information about each instrument used to measure the actual cue values of the performance, student, and task cues. Additional information on some instruments is provided here. To assess the quality of instruments measuring knowledge and understanding (e.g., general reading comprehension, reproduction test, prior knowledge), we used three quality indicators: question difficulty, discrimination, and reliability (Van Berkel & Bax, 2006; Van den Brink & Mellenbergh, 1998). If an instrument performed below par on ≥2 indicators, we excluded the variable.

Performance Cue Values – Diagram Cues. We coded students' diagrams to measure diagram cues, using an existing answer format (cf. Author, 2014; 2019). First, the facts in the diagrams were coded. The answer format contained a list of facts and facts pertained to details in the text that were not essential for understanding the cause-and-effect relations. Each fact was assigned 0 (incorrect/not mentioned in the text) or 1 (correct). Three assistants coded 60 diagrams (Krippendorff's alpha: 0.99).

Second, we coded diagram elements (i.e., causes/effects). Elements were coded as correct when matching the answer format (0–4 per text) or as commission when an element in a student's

diagram was not in the answer format. Furthermore, we determined omissions (blank boxes/question marks). Two assistants coded 60 diagrams (Krippendorff's alpha: 0.96). Fourth, the number of correct relations (i.e., correct combination of two elements; per text: 0–4) was coded (Krippendorff's alpha: 0.91).

Student Cue Values - Students' General Reading Comprehension Level. To measure students' general reading comprehension level, we used a cloze test (cf. Kamalski, 2007). The cloze test consisted of an expository text derived from Author (2014) that was comparable in length and difficulty to the main texts of this study. In the text (215 words), 20 words were omitted and students had to complete missing words. The test was piloted and items that were too easy were replaced. The item difficulty varied and the test was not too easy or too difficult; the percentage of students that answered an item correctly ranged from 9.7% to 92% with an average of 62% ($SD = 22\%$; cf. Van Berkel & Bax, 2006; Van den Brink & Mellenbergh, 1998). The item-rest correlations of all items were sufficient for 14 of the 20 items ($M = 0.18$; $SD = 0.08$), indicating that the items discriminated well between students with low and high test scores (cf. Van Berkel & Bax, 2006; Van den Brink & Mellenbergh, 1998).

Student Cue Values - Students' Ability to Reproduce Facts. We used a text and test items for measuring students' retention of facts (Author, 2014). Two assistants coded 90 students' answers (Krippendorff's alpha: 0.98). The item difficulty varied and the test was not too easy or too difficult; the percentage of students that answered an item correctly ranged from 18.8% to 76.8% with an average of 51% ($SD = 27\%$). The item-rest correlations of all items were sufficient for four of the five items ($M = 0.17$; $SD = 0.02$).

Student Cue Values - Students' IQ. Although the shortened Raven Progressive Matrices test showed high internal consistency in previous research (Bilker et al., 2012), Omega was moderate in our sample (0.54). Overall, the item difficulty varied ($M_{\text{proportion correct adjusted for chance}} = 0.63$, $SD = 0.23$, range = 0.19 - 0.86) and items were not too difficult or too easy given that all items had p-values above chance. The item-rest correlations of all items were sufficient to very good ($M = 0.27$; $SD = 0.07$).

7. Teacher measures

7.1. Judgments of students' text comprehension

Per text, teachers indicated how many points they thought each student scored (0–8). The information they could use were: students' completed diagrams (performance cues), information they knew about their student (student cues), information they remembered about the expository texts and test (task cues).

7.2. Cue-Utilization

Teachers were asked 'What did you base your judgment upon? Please be as complete as possible'. They received a cue-list (Table 1, columns 1–3) and the experimenter explained each cue. Additionally, an explanation of the meaning and measurement of each cue was available (not printed in Table 1) to ensure teachers interpreted the cues as intended.

We piloted the cue-list with two teachers resulting in a cue-list of 28 items. That is, additional to the cues in Table 1, four cues were originally present on the cue-list but were omitted from our analyses. Two cues appeared redundant (text difficulty and readability), one turned out to be impossible to score reliably (spelling/grammatical mistakes), and the instrument to determine the actual cue value of prior knowledge was insufficient regarding all three quality indicators.

The order of the cue types (i.e., student, performance, and task cues) on the cue-list was systematically varied using a Latin-square design, resulting in six versions. There were no significant differences in cue-utilization, judgment height, and monitoring accuracy between versions (all p's < 0.05). For each judgment, teachers indicated which cue(s) they used (0 = not used; 1 = used).

Finally, to check whether using a cue-list did not affect teachers' cue-utilization, we compared teachers' cue-utilization to teachers' cue-utilization in a previous study using a think-aloud procedure without a cue-list (Author, 2020). In Author (2020), teachers mostly focused on the completeness (e.g., do diagrams contain all necessary elements) and correctness of the diagrams (elements and relations). As this is highly similar to our results, there does not seem to be a reason to assume that providing teachers with a cue-list affected their cue-utilization.

7.3. Cue judgments

In the judgment + cue-list + cue-value-judgment condition, teachers also made judgments for used cues. If a teacher for example indicated that they used student interest and IQ, they were asked to answer the interest scale for this student and to indicate how many questions of the Raven standard progressive matrices the student answered correctly. For all student cues for which we used self-report scales (e.g., student interest) or student tests (e.g., general reading comprehension level), teachers viewed the questions of the scales/test and if relevant (e.g., on the Raven test), correct answers. The minimum/maximum cue judgment values corresponded to the minimum/maximum of the instruments used to measure cue values. For cues for which the actual cue values were obvious, teachers did not estimate the cues (i.e., student's gender, omissions, text length and position, time to complete diagram, and mean number of words in the diagram boxes). For learning problems, teachers indicated whether students had (1) or did not have a learning problem.

8. Procedure

8.1. Students

Both sessions took place in a computer room at the participants' school during a lesson period, with the whole class present. Students completed the tasks individually at their own pace on a computer in two sessions (Fig. 2). Although the teacher was present, a researcher led the session and made sure students worked in silence on the tasks. In session two, students practiced reading and diagramming guided by a movie clip. During practice, they read two texts, completed two diagrams, and two test questions. Additionally, they compared their answers to an answer model. The movie clip contained explanation on how the task worked and provided and discussed the answer models.

9. Teachers

The teacher part took place in individual sessions, scheduled after student session 2 was completed. After providing general information, teachers read the students' instructions about the reading tasks and test, including example test questions (Fig. 3). Teachers read the three texts and judged students' text comprehension. After having made judgments for each student and each of the three texts, teachers gave restudy rankings, indicating in what order each student should restudy the texts and indicated how they thought the students judged their own test score.[3] Teachers were

[3] Restudy selections and other judgments fall outside the scope of this article.

Activities in Students' Session 1 and 2

	Student session 1 (±50 min)	Student session 2 (±50 min)
1	Instruction movie clip	Instruction movie clip + practice diagramming
2	General reading comprehension test	Read three expository texts + indication of no. of words they found difficult per text + completion of interest scale per text.
3	Provide general information (gender, birthdate, country of birth [self/mother/father], languages spoken at home, learning problems, and grades)	Complete a pre-structured diagram about each text on paper (without seeing the texts).
4	Prior knowledge test	Complete test questions about the texts
5	General effort scale	Complete other measures not included in the current study (judge own and peers' understanding)
6	Read text for reproduction test	
7	Big five scales	
8	Take reproduction test	
9	Complete other measures not included in the current study (e.g., liking of peers, familiarity with peers)	

Fig. 2. Activities in students' session 1 and 2.

asked to think out loud while making judgments. In all conditions, teachers were provided with information from which they could deduce cues: 1) students' completed diagrams about the texts (performance cues), 2) students' names (student cues), and (3) the task (task cues). Teachers first made all judgments in the judgment-only condition, then in the judgment + cue-list condition, and finally judgment + cue-list + cue-value-judgment condition to prevent carry-over effects. Within each condition, teachers always started with a 'practice student' to get familiar with the procedure of the condition. In the judgment-only condition, they practiced the procedure with the practice student for all three texts. That is, they made judgments about students' text comprehension for each of the three texts and then they made a restudy decision. In the judgment + cue-list condition and the judgment + cue-list + cue-value-judgment condition, teachers practiced the procedure with a practice student, but only for one text because of time constraints. Because they only practiced with one text, they did not make restudy decisions for the practice student in these two conditions. In the judgment + cue-list condition, teachers − in addition to making judgments− also indicated which cues they had used for their monitoring judgments (see Fig. 3). In the judgment + cue-list + cue-value-judgment condition, teachers − in addition to making judgments and indicating cue-use − also judged the values of the used cues. Data of the practice students was not included in the analyses.

10. Used indices and analyses

10.1. Monitoring accuracy

We used bias and absolute accuracy as indices of teachers' monitoring accuracy. Bias was calculated by subtracting a student's test score from a teacher's judgment. Scores range from −8 to +8; scores closer to zero indicate more accurate judgments, negative scores indicate underestimation, positive scores overestimation. Absolute accuracy is the absolute difference between teachers'

Activities Teacher Session

1. Demographics (birthdate, gender, birth country [self/mother/father], education, experience)
2. View instructions that the students have had
3. Read the three texts
Judgment-only condition
4. Make monitoring judgments for practice student for text 1, 2, and 3 successively
5. Make restudy decisions for practice student
6. Make monitoring judgments for student 1 for text 1, 2, and 3 successively
7. Make restudy decisions for student 1
8. Repeat steps 6 and 7 for students 2 and 3
Judgment + cue-list condition
9. Make monitoring judgment for practice student for text 1
10. Complete cue-list for monitoring judgment for practice student for text 1
11. Make monitoring judgment for student 4 for text 1
12. Complete cue-list for monitoring judgment student 4 for text 1
13. Repeat steps 11 and 12 for texts 2 and 3
14. Make restudy decision for student 4
15. Repeat steps 11-14 for students 5 and 6
Judgment+cue-list+cue-value-judgment condition
16. Make monitoring judgment for practice student for text 2
17. Complete cue-list for monitoring judgment for practice student for text 2
18. Judge cue-values for used cues for practice student for text 2
19. Make monitoring judgment for student 7 for text 1
20. Complete cue-list for monitoring judgment for student 7 for text 1
21. Judge cue-values for used cues for student 7 for text 1
22. Repeat steps 19-21 for texts 2 and 3
23. Repeat steps 19-22 for students 8 and 9

Fig. 3. Activities teacher session.

judgments and students' test scores. Scores range from 0 to +8; scores closer to zero indicate more accurate judgments.

10.2. Cue-diagnosticity

To measure cue-diagnosticity, we calculated correlations between actual cue values (cf. Table 1 for instruments) and students' test scores. For those cues for which a negative value meant high diagnosticity (i.e., omissions and commissions, number of difficult words in a text, learning problem, text position and length), we used the unsigned correlation. A cue was highly diagnostic when cue values highly correlated to students' test scores.

10.3. Used-cue value judgment accuracy

In the judgment + cue-list + cue-value-judgment condition teachers were also asked to judge the values of the cues used. We calculated the cue-value judgment accuracy in terms of bias; for this we subtracted the actual cue value from the judged cue value. So if a teacher for example judged that a student had two correct facts in their diagram, whereas this student had, in reality, five correct facts, the bias score was $2-5=-3$, meaning that the teacher

underestimated the number of correct facts. We also calculated absolute judgment accuracy by calculating the absolute difference between teachers' judgments of cue value and actual cue values. So for the aforementioned example, the absolute cue-value judgment accuracy would be three (5−2). Because scales differed per cue (Table 1), the range of used-cue value judgment accuracy varied per cue. In our analyses, we used z-scores.

10.4. Analyses

For RQ1 (cue-diagnosticity), we provide correlations between actual cue values and students' test scores. Regarding RQ2 (cue-utilization), we provide descriptives and occurrences of cue(s) used for single judgments (i.e., cue-use pattern). We restricted the description to those cues-use patterns that were used in ≥10% of the judgments. For RQ4 (relation cue-utilization and used-cue value judgment accuracy and monitoring accuracy), we used multilevel analysis (judgment (level1), student (level 2), teacher (level 3)). Teachers only judged cues that they had used; therefore there were many 'missing' values for the judgments of those cues that were not used. For some cues, cue judgments were missing for as many as 97.9% of the cases (e.g., student's nationality; this cue was thus seldomly used). For the used-cue value judgment accuracy model, we only selected cues that had less than 60% missing values (cf. Table 3). For the cue-utilization model, we included all cues.

11. Results

Generally, teachers overestimated students' test scores with 1.15 points and their judgments deviated, in an absolute sense, 2.19 points on average from students' actual test scores (cf. Table 2).

11.1. Cue-diagnosticity (RQ1)

As expected, performance cues were, on average, more diagnostic than student and task cues (Table 3). Task cues had the lowest diagnosticity. Yet, within cue categories, we saw substantive variation. As expected, the performance cue 'number of correct facts[4]' was hardly diagnostic (0.08), whereas the 'number of correct relations' was highly diagnostic (0.59). Another highly diagnostic cue was the 'number of correct elements' (0.63). The cue 'omissions' was somewhat less diagnostic but still moderately to strongly correlated to students' test scores (0.45). All student cues had low diagnosticity (all < 0.30). Within this category, 'general reading comprehension level' (0.25) and 'IQ' (0.22) were relatively most diagnostic. All task cues had low diagnosticity.

11.2. Cue-utilization (RQ2)

Per judgment, teachers used on average 6.35 cues ($SD = 3.94$) with a minimum of 1 and maximum of 24 (out of 28) cues. On average, they used 3.19 diagram cues ($SD = 1.77$), 2.25 student cues ($SD = 2.30$), and 0.92 task cues ($SD = 1.33$) per judgment. In many of their judgments, teachers used cues that were highly diagnostic. For example, they used correct elements, relations, and omissions in over 50% and students' general reading comprehension level and IQ in over one third of their judgments. Yet, two low diagnostic cues (i.e., correct facts and students' effort) were also used relatively often; in about two third and one third of the judgments, respectively. Differences in cue-utilization were small between conditions (see Supplemental material).

For the total of 813 judgments we encountered 456 unique cue-use patterns, occurring between 1 and 28 times.[5] The patterns occurring >10 are reported in Table 4. The fact that the most common pattern −i.e., omissions, correct facts, elements, and relations− was only used 28 times (in 813 judgments; i.e., in 3.44%), indicates that there was not a single cue-use pattern that stood out. Seven out of eight cue-use patterns in Table 4 consisted of performance cues only and mostly included correct relations, facts, and elements (6 out of 8 patterns). Furthermore, teachers sometimes only used one cue, that is, omissions.

11.3. Teachers' used-cue value judgment accuracy (RQ3)

Teachers mostly struggled with accurately judging performance cues; their judgments deviated on average around 30% from actual cue values. They mostly *overestimated* cue values (Table 3). Correct elements, for example, which was a highly diagnostic cue, was overestimated by about 30%. Correct facts (low diagnosticity) was overestimated by about 50%. As for student cues, teachers' used-cue value judgment accuracy differed between cues. For some student cues, judgments were remarkably accurate (e.g., conscientiousness: 0.29%; grades other subject: 3.3%; student's interest: 0.33%) whereas for other student cues, judgments were quite inaccurate (e.g., students' ability to reproduce facts: overestimation of 33%; student's nationality: teachers thought the student and/or their parents were non-Dutch whereas they were [28.8%]). Teachers judged the number of facts in the text (task cue) relatively accurate (2.67% deviation) but overestimated the number of difficult words in the text by about 24%.

11.4. Cue-utilization and used-cue value judgment accuracy vs. monitoring accuracy (RQ4)

The majority of variance in teachers' monitoring accuracy of students' test scores was situated at the judgment level (bias: 73%; deviation: 81%). Smaller parts of the variance resided at the teacher (bias: 11%; deviation; 3%) and student level (bias: 16%; deviation: 16%). Monitoring accuracy thus mainly varied from judgment to judgment.

When teachers used omissions as a cue, their monitoring accuracy (deviation) was higher (Table 5). In contrast, using students' general reading comprehension levels, grades for other subjects, nationality, extraversion, and IQ was related to more overestimation (bias). When teachers judged the correct relations in students' diagrams and students' general effort levels in class more accurately (deviation), their monitoring was more accurate (deviation and bias; Table 5).

12. Discussion

We investigated teachers' monitoring accuracy of students' text comprehension. Students completed pre-structured diagrams representing causal relations in the texts they had read. While judging students' text comprehension (i.e., test performance), teachers had access to these diagrams (giving access to performance cues such as correct relations and omissions in students' diagrams), and to students' names (giving access to student cues such as IQ and gender). They had also read the texts and seen example test questions beforehand (giving access to task cues such as text length and text position). We explored how diagnostic a

[4] The number of correct facts refers to elements that are not essential for the causal relations and that were thus not part of the test.

[5] We restricted ourselves to cues that were used in ≥10% of the judgments (cf. Table 3).

Table 3
Cue-diagnosticity, teachers' self-reported cue-utilization, actual cue values, teachers' cue judgments and teachers' used-cue value judgment accuracy per cue.

		Cue-diagnosticity	Cue-utilization	Actual cue values		Cue value judgment		Used-cue value judgment accuracy					
									Deviation		Bias		% deviation[a]
	Min/max scales												
		M	M	SD	M	SD	M	SD	M	SD	M	SD	
Performance cues													
Diagram – commission errors	0–4	.29	.24	.43	.31	.71	.87	.95	.72	.82	-.35	1.03	14%
Diagram – correct facts	0–5	.08	.63	.48	.16	.40	2.74	1.25	2.60	1.30	2.59	1.31	51.6%
Diagram – correct elements	0–6	.63	.58	.49	.84	1.36	2.48	1.34	.72	.87	.08	1.13	27.33%
Diagram – correct relations	0–4	.59	.67	.47	1.96	1.39	2.43	1.40	.98	.99	.51	1.29	30.99%
Diagram – extensiveness formulations	0-∞	.38	.32	.47	4.27	1.77	Cue not judged by the teacher						
Diagram – omission errors	0–4	.45	.55	.50	.58	1.07	Cue not judged by the teacher						
Diagram – time (min) to complete	0-∞	.03	.16	.37	2.30	1.54	Cue not judged by the teacher						
Mean performance cues		**.40**	**.45**	**.46**					**1.26**	**1.00**	**.71**	**1.24**	**30.98%**[b]
Student cues													
Student - conscientiousness	0–7	.04	.26	.44	4.49	1.12	4.47	1.27	.97	.88	.14	1.30	−0.29%
Student - effort	0–4	.08	.31	.46	2.94	.56	2.55	.40	.56	.37	-.30	.60	−9.75%
Student - extraversion	0–7	-.04	.06	.24	4.92	1.29	4.32	1.34	2.00	1.48	.19	2.51	−8.57%
Student - fact reproduction	0–5	.16	.08	.27	2.68	1.13	4.33	.71	2.14	1.20	1.86	1.61	33%
Student - gender	1–2	-.08	.03	.16	1.45	.50	Cue not judged by the teacher						
Student - grade other subjects	0–10	.19	.06	.25	5.00	.00	5.33	1.32	3.43	3.31			3.3%
Student - grade subject teacher	0–10	.03	.28	.45	6.48	.67	6.62	1.13	.45	.05	.05	.49	1.4%
Student - interest in text topic	1–4	.19	.19	.39	2.37	.71	2.35	.70	.73	.63	.02	.97	−0.33%
Student - IQ	0–9	.22	.32	.47	6.41	1.72	6.71	2.37	1.96	1.59	-.08	2.53	3.33%
Student - learning problems[c]	0–1	.07	.06	.23	.10	.29	.06	.23	.90	.30	.29	.91	−84%
Student - nationality	0–5	.14	.02	.14	4.77	.85	3.33	.50	3.33	.50	.11	3.22	−28.8%
Student - general reading comprehension level	0–20	.25	.31	.46	12.65	2.86	13.62	3.95	3.68	2.32	1.30	4.16	5.38%
Student - self-efficay	0–4	.10	.16	.37	2.83	.66	2.71	.63	.70	.50	-.39	.76	−3%
Mean student cues		**.10**	**.14**	**.29**					**1.41**	**1.00**	**.28**	**1.69**	**15%**[b]
Task cues													
Task - difficult words in text	0-∞	.12	.11	.31	1.20	1.25	5.95	4.70	4.55	4.58	4.46	4.66	23.75%
Task - facts in text	5–7	.03	.22	.41	6.00	.82	6.08	3.17	2.70	1.74	.11	3.22	2.67%
Task - text length (no. of lines)	12–14	.03	.07	.25	13.00	.82	Cue not judged by the teacher						
Task - text position	1–3	.03	.08	.27	2.00	.82	Cue not judged by the teacher						
Mean task cues		**.05**	**.48**	**.31**					**3.63**	**3.16**	**2.29**	**3.94**	**13.21%**[b]

Note. Cue-diagnosticity: min = −1 meaning low diagnosticity, max = +1 meaning high diagnosticity; cue-utilization: min = 0, max = 1; used-cue value judgment accuracy: closer to 0 is more accurate. Cue-utilization is coded as 0 (not used) or 1 (used); the mean indicates proportion of judgments for which the particular cue is used.

[a] Calculated as: ((cue judgment − actual cue value)/nr of scale points)*100. A positive value indicates that a teacher's overestimation of the cue value and a negative value indicates underestimation. If the max for a cue was ∞, we used the maximum of the teachers' cue judgment.

[b] Mean percentage in absolute sense.

[c] To calculate teachers' used-cue value judgment accuracy for learning problems, we only considered the combination of a teacher who indicated that (s)he used this cue (score = 1) with that student actually having a learning problem (score = 1) as accurate. Cases in which the teacher did not use it and the student did not have it were not counted as accurate because not using it was the default value for this teacher variable; we did not ask the teacher to explicitly judge whether or not the student had each learning problem, we only asked whether they used it.

Table 4
Cue-use patterns occurring >10 times.

Cue-use pattern *(total no. of patterns: 456; total no. of judgments: 813)*	No. times the cue-use pattern occurs
omission errors(p)/facts(p)/elements(p)/relations(p)	28
facts(p)/elements(p)/relations(p)	28
omission errors(p)/facts(p)/relations(p)	21
omission errors(p)/commissions(p)/facts(p)/elements(p)/relations(p)/extensiveness(p)	20
omission errors(p)/commision(p)/facts(p)/elements(p)/relations(p)	16
omission errors(p)/commision(p)/facts(p)/elements(p)/relations(p)/diffwordstext(t)	16
omission errors(p)	16
fact(p)/elements(p)/relations(p)/extensiveness(p)	13

Note. (p) = performance cue, (t) = task cue.

wide range of performance, student, and task cues were for students' text comprehension (RQ1), what patterns in teachers' cue-utilization could be observed (RQ2), and how accurately teachers could judge the values of the cues they had used (used-cue value

Table 5
Model results for multilevel models of teachers' judgment accuracy of students' test scores predicted by cue-utilization and used-cue value judgment accuracy (unstandardized coefficients).

		Cue-utilization				Used-Cue Value Judgment Accuracy							
		Deviation		Bias		Dev/dev[a]		Bias/bias[b]		Dev/bias[c]		Bias/dev[d]	
		b (SE)	*p*	*b (SE)*	*p*	*b (SE)*	*p*	*b (SE)*	*p*	*b (SE)*	*p*	*b (SE)*	*p*
	Intercept	2.33 (.24)	.00	.81 (.42)	.05	2.58 (.23)	.00	1.99 (.25)	.00	2.25 (.27)		2.36 (.21)	.00
	Commission errors	-.14 (.18)	*ns*	-.33 (.27)	*ns*	X		X		X		X	
	Correct facts	-.07 (.19)	*ns*	.45 (.33)	*ns*	.11 (.14)	*ns*	.03 (.29)	*ns*	-.24 (.24)	*ns*	.39 (.24)	*ns*
	Correct elements	-.06 (.19)	*ns*	-.11 (.35)	*ns*	.14 (.29)	*ns*	.68 (.44)	*ns*	-.30 (.41)	*ns*	.35 (.34)	*ns*
	Correct relations	0.12 (.19)	*ns*	-.10 (.31)	*ns*	1.60 (.23)	**.00**	.92 (.65)	*ns*	2.69 (.46)	**.00**	.56 (.48)	*ns*
	Extensiveness formulations	.28 (.16)	*ns*	.06 (.29)	*ns*	X		X		X		X	
	Omission errors	-.42 (.18)	***.02***	.01 (.31)	*ns*	X		X		X		X	
	Time to complete diagram	-.12 (.23)	*ns*	.01 (.34)	*ns*	X		X		X		X	
Student characteristics	Conscientiousness	-.13 (.17)	*ns*	-.34 (.26)	*ns*	X		X	*ns*	X		X	*ns*
	Effort	.06 (.18)	*ns*	.07 (.29)	*ns*	.57 (.23)	**.01**	-.24 (.34)	*ns*	.81 (.23)	**.00**	-.34 (.27)	*ns*
	Extraversion	-.33 (.28)	*ns*	−1.04 (.50)	**.04**	X		X		X		X	
	Fact reproduction	.32 (.34)	*ns*	.35 (.42)	*ns*	X		X		X		X	
	Gender	-.01 (.64)	*ns*	.57 (.99)	*ns*	X		X		X		X	
	Grade other subject	.27 (.30)	*ns*	1.19 (.54)	**.03**	X		X		X		X	
	Grade subject teacher	-.14 (.17)	*ns*	-.35 (.26)	*ns*	X		X		X		X	
	Interest	.12 (.14)	*ns*	.09 (.36)	*ns*	X		X		X		X	
	IQ	.19 (.18)	*ns*	.61 (.25)	**.01**	X		X		X		X	
	Learning problems	.39 (.39)	*ns*	.74 (.60)	*ns*	X		X		X		X	
	Nationality	.77 (.41)	*ns*	1.80 (.60)	**.003**	X		X		X		X	
	Reading comprehension	.03 (.16)	*ns*	.60 (.28)	**.03**	-.04 (.23)	*ns*	-.39 (.44)	*ns*	-.43 (.25)	*ns*	-.16 (.30)	*ns*
	Self-efficacy	.00 (.18)	*ns*	.21 (.34)	*ns*	X		X		X		X	
Task characteristics	Difficult words in the text	.32 (.22)	*ns*	-.56 (.53)	*ns*	X		X		X		X	
	Facts in the text	-.20 (.16)	*ns*	-.43 (.29)	*ns*	X		X		X		X	
	Text length	-.04 (.27)	*ns*	.22 (.39)	*ns*	X		X		X		X	
	Text position	-.02 (.32)	*ns*	-.09 (.37)	*ns*	X		X		X		X	
	Residual variance	3.29 (.22)	.00	6.50 (.39)	.00	1.07 (.23)	.00	2.79 (.56)	.00	1.59 (.31)	.00	1.57 (.36)	.00
	R^2	.04 (.02)		.08 (.03)		.54 (.10)		.32 (.17)		.61 (.11)		.30 (.18)	

Note. For the used-cue value judgment accuracy model, we only selected cues that had less than 60% missing values.
[a] Dev/dev = used-cue value judgment accuracy deviation score (IV) and judgment accuracy of students' text comprehension deviation score (DV).
[b] Bias/bias = used-cue value judgment accuracy bias score (IV) and judgment accuracy of students' text comprehension bias score (DV).
[c] Dev/bias = used-cue value judgment accuracy deviation score (IV) and judgment accuracy of students' text comprehension bias score (DV).
[d] Bias/dev = used-cue value judgment accuracy bias score (IV) and judgment accuracy of students' text comprehension deviation score (DV).

judgment accuracy; RQ3). The main aim was to investigate to what extent teachers' cue-utilization and the degree to which they accurately judged the values of the used cues was related to teachers' monitoring accuracy of students' text comprehension (RQ4). Our findings show that teachers generally overestimated students' test performance. Their monitoring accuracy was higher when teachers ignored non-diagnostic cues and used diagnostic cues —but only when they were able to accurately assess the value of those diagnostic cues.

12.1. Cue-diagnosticity (RQ1)

Monitoring accuracy is considered to depend on how diagnostic used cues are, that is, how predictive they are of test performance (Koriat, 1997). However, cue-diagnosticity is often not measured. By measuring actual cue-values we could determine cue-diagnosticity. Overall, performance cues were most diagnostic, then student cues, followed by task cues. As expected, the number of correct relations in students' diagrams was highly diagnostic of students' test scores (cf. Author, 2014; 2020). Correct elements and omissions in students' diagrams were moderately to highly diagnostic. Importantly, not all performance cues were diagnostic; as expected, correct facts in students' diagrams, which was used in many teachers' judgments, had low diagnosticity as did commissions in students' diagrams. All student and task cues had low diagnosticity. These findings substantiate the widely held assumption that performance cues are highly diagnostic, and more diagnostic than student and task cues. However, the variability in the diagnosticity of performance cues shows that caution is needed when designing interventions to improve teachers' monitoring accuracy. Only the use of certain performance cues (here: relations, elements, and omissions) should be promoted, based on their actual diagnosticity for the to-be-judged task.

12.2. Cue-utilization (RQ2)

To gain more insight in the judgment process, we investigated the number, type, and patterns of cues used. The number of cues used and the extent to which each cue-type was used, was similar to findings of Author (2020), p. 6.35 cues were used on average per judgment and teachers mostly used performance cues, then student and then task cues. The cues with the highest diagnosticity (correct elements, relations, and omissions) were used in the majority of judgments. Yet, teachers also used performance and student cues with low diagnosticity (i.e., facts in students' diagrams, students' effort in class, grades for the teacher's subject, general reading comprehension level, IQ) to a considerable extent, even though they were made aware that they had to judge students' test scores and that the test was about text elements and relations. We found as many as 456 unique cue patterns on a total of 813 judgments and there was not a single pattern that stood out for being used often. However, the most frequently used patterns only or mainly contained performance cues.

These findings show that teachers draw upon quite some information when making judgments, including non-diagnostic information. Future research could investigate whether teachers' monitoring accuracy would improve from encouraging them to limit the number of cues they use and focus on diagnostic

performance cues.

12.3. Used-cue value judgment accuracy (RQ3)

For accurate monitoring, focusing on diagnostic cues and ignoring non-diagnostic cues may be a necessary but not sufficient condition: Teachers should also accurately judge the value of the used (e.g., judge how many relations students completed correctly in their diagram). Teachers' judgments of performance cues – which had the highest diagnosticity – appeared to be least accurate; teachers, on average, overestimated these cue values by 30%. Two highly diagnostic cues (correct relations and elements) were, respectively, overestimated by 31% and 27%. This overestimation is in line with what we generally see in the literature about teachers' judgments of students' achievement (Südkamp et al., 2012; Urhahne & Wijnia, 2021). A possible explanation for this may be that teachers did not use the same standards as we in deciding whether relations or elements was correct. Yet, the correct answers were rather straightforward as the texts contained the correct elements and relations and the teachers knew the texts. Perhaps, teachers suffered from the leniency effect as suggested by Urhahne and Wijnia (2021). That is, teachers may "not take sufficient account of factors such as students' forgetting of subject matter, limited testing time, lack of effort, excitement, and test anxiety (Hosenfeld et al., 2002)." (Urhahne & Wijnia, 2021, p. 6). Therefore, even when particular cues are easy to judge, other factors may still distort teachers' judgments. In addition to not taking into account particular factors, teachers may also have taken non-diagnostic student cues into account when judging students' diagrams, which may also have hampered their cue judgment accuracy.

Cue-Utilization, Used-Cue Value Judgment Accuracy and Monitoring Accuracy (RQ4).

Merely *using* highly diagnostic cues was insufficient for accurate monitoring; there was no effect of using either of the two most diagnostic cues on teachers' monitoring accuracy. Yet, when teachers *judged* one of these most diagnostic cues (i.e., correct relations in students' diagrams) *more accurately* when using it, their monitoring of students' text comprehension was also more accurate. It may seem self-evident that when relations in students' diagrams are judged more accurately, students' test scores are also judged more accurately as the test focuses on students' understanding of relations. Yet, the relation between used-cue judgment accuracy and monitoring accuracy of students' performance has not been investigated before.

Furthermore, we found that using some of the low diagnostic cues *hampered* teachers' monitoring accuracy (i.e., students' general reading comprehension levels, grades for other subjects, nationality, extraversion, IQ). A similar effect was found in Author (2018) when using a problem-solving task in Mathematics: teachers' monitoring of students' mathematics achievement was less accurate when they had non-diagnostic student cues available in addition to diagnostic performance cues. Teachers in our study judged the low diagnostic cues quite accurately (exception: students' nationality). Finally, for one cue (i.e., omissions), mere usage was related to more accurate monitoring. Yet, judgment of this cue was hardly needed as it only involved counting the number of blank boxes and question marks in diagrams. Surprisingly, when teachers judged the non-diagnostic cue students' general effort in class more accurately, their monitoring was also more accurate whereas mere use of this cue did not foster monitoring accuracy. For those effort judgments that were very accurate (absolute deviation < 0.30), the mean level of students' effort was somewhat lower (2.6) than for those effort judgments that were more inaccurate (absolute deviation >1) in which case the mean was 3.3. Perhaps, when monitoring effort more accurately and when student effort was relatively low, teachers may have lowered their judgments of students' test scores based on the somewhat lower effort level. Given that teachers generally overestimated students' test scores, lowering their judgments may have resulted in more accurate judgments of students' test scores. Yet, future research should further investigate this tentative explanation.

12.4. Limitations and future research

One limitation is that we measured cue-diagnosticity by calculating overall correlations between actual cue values and students' test scores. This group-level diagnosticity is useful when e.g., designing interventions. Nevertheless, it may be that a particular cue is somewhat more diagnostic for one student than for another student.

Furthermore, the instruments for measuring actual cue values of students' IQ, ability to reproduce facts, and general reading comprehension level did not perform sufficiently on one of the three quality indicators (i.e., internal consistency). We therefore need to interpret these results with caution. The low internal consistency may make it harder for teachers to judge these cues given that answers on items within cues are not necessarily consistent. Nevertheless, teachers judged the actual cue values of students' IQ and general reading comprehension very accurately (deviation 3–5%). Future research could investigate whether teachers' judgments of these cues would be similarly accurate when using instruments with higher internal consistency.

In addition, differences between texts regarding, for instance, length and difficulty were small. This may have caused low diagnosticity and may have prevented teachers – if they were aware of this – from making (more) use of these cues. Future research could further investigate the diagnosticity and cue-utilization of task cues when there is more variation in task characteristics. Moreover, although findings from RQ4 are highly relevant, our data do show whether the beneficial effect of accurately judging diagnostic cues occurred because teachers only used diagnostic cues, judged these cues accurately, and ignored non-diagnostic cues, or whether they did also used non-diagnostic cues but using these did not hamper their monitoring accuracy when using and accurately judging diagnostic cues. Future research could investigate this issue further.

Finally, we focused on teachers' monitoring of students' text comprehension. In other domains and with other tasks, effects of teachers' cue-utilization and used-cue judgment accuracy on their monitoring accuracy could be different. Yet, a previous study has found that when monitoring problem-solving tasks in Mathematics, teachers were most accurate when they only had diagnostic performance cues available (using anonymized student work) compared to having only student cues or performance and student cues (Author, 2018). Thus, similar to our findings, using non-diagnostic student cues seems to hamper teachers' monitoring accuracy also in other domains with other tasks, such as Mathematics.

13. Conclusion

The current study addresses teachers' monitoring of students' text comprehension when learning from texts describing causal relations, which is relevant for most subjects in secondary education. Prior research has shown that making information containing diagnostic information about students' text comprehension may be insufficient to improve teachers' monitoring accuracy. Our findings show that teachers also need to ignore non-diagnostic cues. Importantly, this study shows that deducing diagnostic cues from available information is a necessary but not sufficient condition for higher monitoring accuracy. Rather, teachers also need to judge cue

values accurately if they are to accurately monitor students' text comprehension. Thus, although it has hardly received attention in the literature, teachers' used-cue value judgment accuracy seems to form an indispensable part of the monitoring process. If future research would show this finding to be robust, it could add significantly to theoretical and/or process models of teacher monitoring such as the cue-utilization model.

Our findings also have relevance for designing interventions to improve teachers' monitoring accuracy. For instance, it may be useful to raise teachers' awareness of which cues are diagnostic (and should be used) and which are not (and should be ignored) and to help teachers in accurately monitoring the most diagnostic cues either by themselves or with the aid of technology such as learning analytics.

Author note

Correspondence concerning this manuscript should be addressed to Janneke van de Pol (j.e.vandepol@uu.nl), Utrecht University, Department of Education, PO Box 80.140, 3508 TC Utrecht, The Netherlands, +31 302531796. Preliminary results of this study have been presented at the biennial conference of the European Association for Learning and Instruction, Aachen, Germany, 2019.

Acknowledgments

We would like to thank Marloes Berkers, Jonne Bloem, Luca Clercx, and Nynke Heegstra, who collected the data and coded the students' tests and diagrams and Mirjam Moerbeek with help on the power analysis. This work was supported by the Netherlands Organization for Scientific Research (grant number: 451-16-012).

Appendix A

Instructions and example test question

Text reading

Students received the following general instructions: "You're about to read several texts. Try to understand these as good as you can! You can only read each text once. When you have read a text, please continue with the next one. You cannot go back."

Per text, they received the following instructions: "Please read this text carefully. You cannot look back in the text when you will complete diagrams and take the test."

Diagrams

Instructions: "Please complete the diagram for the text [title text] that you have read. If you are unable to complete a box, please fill out a ?".

Test

Instructions: "On this test, you will get questions about causes and effects in the texts and the order between these causes and effects. Try to answer each question. Good luck!"

Example test question about "concrete constructions"

Elevators in concrete buildings often need to be renovated. What are four causes of why these elevators need to be renovated? Provide an answer that is as complete as possible. Also indicate the order of the four causes, for example by using the words: and, therefore, because, because of that, for that reason, for those two reasons, first, second, this has two consequences. Also use the following sentence in your answer: "renovation of elevators is often needed in concrete buildings".

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tate.2021.103386.

References

Artelt, C., & Rausch, T. (2014). Accuracy of teacher judgments: When and for what reasons?. In *Teachers' professional development* (pp. 27–43). Brill Sense.

Author. (2010). *A; 2010B; 2011; 2014; 2015; 2019A; 2019B; 2020*.

Bennett, R. E., Gottesman, R. L., Rock, D. A., & Cerullo, F. (1993). Influence of behavior perceptions and gender on teachers' judgments of students' academic skill. *Journal of Educational Psychology, 85*(2), 347–356. https://doi.org/10.1037/0022-0663.85.2.347

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

Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation, 13*(5), 401–434. https://doi.org/10.1080/13803610701728311

Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction, 24*, 4–14. https://doi.org/10.1016/j.learninstruc.2012.06.001

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

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

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review, 28*(4), 717–741. https://doi.org/10.1007/s10648-015-9348-9

Funder, D. C. (Ed.). (1999). *Personality judgment: A realistic approach to person perception*. Elsevier.

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

Glock, S., Krolak-Schwerdt, S., Klapproth, F., & Böhmer, M. (2012). Improving teachers' judgments: Accountability affects teachers' tracking decisions. *International Journal of Technology and Inclusive Education, 1*, 89–98.

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

Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition, 37*(7), 1001–1013. https://doi.org/10.3758/MC.37.7.1001

Hecht, S. A., & Greenfield, D. B. (2002). Explaining the predictive accuracy of teacher judgments of their students' reading achievement: The role of gender, classroom behavior, and emergent literacy skills in a longitudinal sample of children exposed to poverty. *Reading and Writing, 15*(7–8), 789–809. https://doi.org/10.1023/A:1020985701556

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

Hoffmann, L., & Böhme, K. (2013). Wie gut können grundschullehrkräfte die schwierigkeit von Deutsch-und mathematikaufgaben beurteilen? Eine untersuchung zur genauigkeit aufgabenbezogener lehrerurteile auf klassenebene [how accurate can elementary school teachers estimate the diffi culty of German and mathematics tasks? An investigation of task related judgment accuracy on class level]. *Psychologie in Erziehung und Unterricht, 61*(1), 42–55. https://doi.org/10.2378/peu2014.art05d

Holder, K., & Kessels, U. (2017). Gender and ethnic stereotypes in student teachers' judgments: A new look from a shifting standards perspective. *Social Psychology of Education, 20*(3), 471–490. https://doi.org/10.1007/s11218-017-9384-z

Hosenfeld, I., Helmke, A., & Schrader, F.-W. (2002). Diagnostische Kompetenz: Unterrichts- und lernrelevante Schülermerkmale und deren Einschätzung durch Lehrkräfte in der Unterrichtsstudie SALVE [Diagnostic competence: How teachers rate the motivational and cognitive characteristics of students in the SALVE study]. *Zeitschrift für Padagogik, 45*, 65–82.

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

Jenkins, L. N., & Demaray, M. K. (2016). Teachers' judgments of the academic achievement of children with and without characteristics of inattention, impulsivity, and hyperactivity. *Contemporary School Psychology, 20*(2), 183–191.

https://doi.org/10.1007/s40688-015-0073-7
Johnston, O., Wildy, H., & Shand, J. (2019). A decade of teacher expectations research 2008–2018: Historical foundations, new developments, and future pathways. *Australian Journal of Education, 63*(1), 44–73. https://doi.org/10.1177/0004944118824420
Kaiser, J., Möller, J., Helm, F., & Kunter, M. (2015). Das schülerinventar: Welche Schülermerkmale die Leistungsurteile von Lehrkräften beeinflussen. *Zeitschrift Fur Erziehungswissenschaft, 18*(2), 279–302. https://doi.org/10.1007/s11618-015-0619-5
Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. *Learning And Instruction, 28*, 73–84. https://doi.org/10.1016/j.learninstruc.2013.06.001
Kamalski, J. M. H. (2007). Coherence marking, comprehension and persuasion. *On the processing and representation of discourse, 158* (LOT).
Karing, C. (2009). Diagnostische Kompetenz von Grundschul- und Gymnasiallehrkräften im Leistungsbereich und im Bereich Interessen. *Zeitschrift für Padagogische Psychologie, 23*(34), 197–209. https://doi.org/10.1024/1010-0652.23.34.197
Klug, J., Bruder, S., Kelava, A., Spiel, C., & Schmitz, B. (2013). Diagnostic competence of teachers: A process model that accounts for diagnosing learning behavior tested by means of a case scenario. *Teaching and Teacher Education, 30*(1), 38–46. https://doi.org/10.1016/j.tate.2012.10.004
Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349
Kostons, D., & de Koning, B. B. (2017). Does visualization affect monitoring accuracy, restudy choice, and comprehension scores of students in primary education? *Contemporary Educational Psychology, 51*, 1–10. https://doi.org/10.1016/j.cedpsych.2017.05.001
Landis, J. R., & Koch, G. G. (1977). *The measurement of observer agreement for categorical data* (pp. 159–174). Biometrics. https://doi.org/10.2307/2529310
Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S. A., & Harackiewicz, J. M. (2010). Measuring situational interest in academic domains. *Educational and Psychological Measurement, 70*(4), 647–671. https://doi.org/10.1177/0013164409355699
Marsh, H. W., Hau, K. T., Artelt, C., Baumert, J., & Peschar, J. L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing, 6*(4), 311–360. https://doi.org/10.1207/s15327574ijt0604_1
McElvany, N., Schroeder, S., Hachfeld, A., Baumert, J., Richter, T., Schnotz, W., & Ullrich, M. (2009). Diagnostische fähigkeiten von Lehrkräften: Bei der Einschätzung von Schülerleistungen und aufgabenschwierigkeiten bei lernmedien mit instruktionalen bildern [diagnostic skills of teachers concerning the evaluation of pupils' achievements and task difficulties using learning media with instructional images]. *Zeitschrift für Padagogische Psychologie, 23*(34), 223–235. https://doi.org/10.1024/1010-0652.23.34.223
Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. *Teaching and Teacher Education, 65*, 48–60. https://doi.org/10.1016/j.tate.2017.02.021
Mizala, A., Martínez, F., & Martínez, S. (2015). Pre-service elementary school teachers' expectations about student performance: How their beliefs are affected by their mathematics anxiety and student's gender. *Teaching and Teacher Education, 50*, 70–78. https://doi.org/10.1016/j.tate.2015.04.006
Moerbeek, M., & Teerenstra, S. (2015). *Power analysis of trials with multilevel data*. Chapman and Hall/CRC.
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
Praetorius, A. K., Berner, V. D., Zeinz, H., Scheunpflug, A., & Dresel, M. (2013). Judgment confidence and judgment accuracy of teachers in judging self-concepts of students. *The Journal of Educational Research, 106*(1), 64–76. https://doi.org/10.1080/00220671.2012.667010
Praetorius, A. K., Koch, T., Scheunpflug, A., Zeinz, H., & Dresel, M. (2017). Identifying determinants of teachers' judgment (in) accuracy regarding students' school-related motivations using a Bayesian cross-classified multi-level model. *Learning and Instruction, 52*, 148–160. https://doi.org/10.1016/j.learninstruc.2017.06.003
Rausch, T., Karing, C., Dörfler, T., & Artelt, C. (2016). Personality similarity between teachers and their students influences teacher judgement of student achievement. *Educational Psychology, 36*(5), 863–878. https://doi.org/10.1080/01443410.2014.998629
Schleinschok, K., Eitel, A., & Scheiter, K. (2017). Do drawing tasks improve monitoring and control during learning from text? *Learning and Instruction, 51*, 10–25. https://doi.org/10.1016/j.learninstruc.2017.02.002
Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762. https://doi.org/10.1037/a0027627
Urhahne, D., & Wijnia, L. (2020). A Review on the accuracy of teacher judgments. *Educational Research Review, 32*, 100374. https://doi.org/10.1016/j.edurev.2020.100374
Van Berkel, H., & Bax, A. (2006). *Toetsen in het hoger onderwijs [Testing in higher education*. Houten/Diegem: Bohn Stafleu Van Loghum.
Van den Brink, W. P., & Mellenbergh, G. J. (1998). *Testleer en testconstructie [Test theory and test construction]*. Amsterdam: Boom.
Weaver, C. A., & Bryant, D. S. (1995). Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text. *Memory & Cognition, 23*(1), 12–22. https://doi.org/10.3758/BF03210553
Webb, M. B. (2015). Exploring the correlation between teachers' mindset and judgment accuracy to reveal the cues behind teachers' expectations. *Doctoral dissertation. Boise, MT: Boise State University*.