Educational Psychology Review (2020) 32:951–977
https://doi.org/10.1007/s10648-020-09560-y

REVIEW ARTICLE

Mapping and Drawing to Improve Students' and Teachers' Monitoring and Regulation of Students' Learning from Text: Current Findings and Future Directions

Janneke van de Pol, et al. *[full author details at the end of the article]*

Published online: 3 August 2020

Abstract
For (facilitating) effective learning from texts, students and teachers need to accurately monitor students' comprehension. Monitoring judgments are accurate when they correspond to students' actual comprehension. Accurate monitoring enables accurate (self-)regulation of the learning process, i.e., making study decisions that are in line with monitoring judgments and/or students' comprehension. Yet, (self-)monitoring accuracy is often poor as the information or cues used are not always diagnostic (i.e., predictive) for students' actual comprehension. Having students engage in generative activities making diagnostic cues available improves monitoring and regulation accuracy. In this review, we focus on generative activities in which text is transformed into visual representations using mapping and drawing (i.e., making diagrams, concept maps, or drawings). This has been shown to improve monitoring and regulation accuracy and is suited for studying cue diagnosticity and cue utilization. First, we review and synthesize findings of studies regarding (1) students' monitoring accuracy, regulation accuracy, learning, cue diagnosticity, and cue utilization; (2) teachers' monitoring and regulation accuracy and cue utilization; and (3) how mapping and drawing affect using effort as a cue during monitoring and regulation, and how this affects monitoring and regulation accuracy. Then, we show how this research offers unique opportunities for future research on advancing measurements of cue diagnosticity and cue utilization and on how effort is used as a cue during monitoring and regulation. Improving measures of cue diagnosticity and cue utilization can provide us with more insight into how students and teachers monitor and regulate students' learning, to help design effective interventions to foster these important skills.

Keywords Self-regulated learning · Student monitoring · Teacher monitoring · Metacomprehension accuracy · Mental effort · Text comprehension

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10648-020-09560-y) contains supplementary material, which is available to authorized users.

Learning from texts is essential for most school subjects, further education, and working life (Biancarosa & Snow, 2004). To learn effectively from texts, students and teachers need to be able to accurately monitor and regulate students' text learning (Dent & Koenka, 2016; Südkamp, Kaiser, & Möller, 2012). Accurate monitoring is important because it directly influences the learning actions students take to self-regulate their learning, such as selecting materials for restudy, allocating study time, seeking help, withdrawing erroneous responses, etc. (Dunlosky & Ariel, 2011; Ghetti, Hembacher, & Coughlin, 2013; Steiner, van Loon, Bayard, and Roebers, 2020). Students and teachers monitor students' text learning well when their judgments about students' text comprehension are accurate; that is, when their judgments are in line with students' actual comprehension as evidenced by test performance (e.g., Griffin, Mielicki, & Wiley, 2019). When regulation decisions such as deciding which text(s) need to be restudied before taking a test are in line with one's (accurate) monitoring judgments and/or text comprehension, regulation is considered accurate (i.e., fitting the students' needs; e.g., Van Loon, De Bruin, Van Gog, Van Merriënboer, & Dunlosky, 2014). In turn, accurate regulation results in better learning outcomes (i.e., better comprehension test performance; Thiede, Anderson, & Therriault, 2003; Thiede, Oswalt, Brendefur, Carney, and Osguthorpe, 2019a; see also Dunlosky & Rawson, 2012). However, students and teachers have difficulties with accurately monitoring students' comprehension, which makes subsequent regulation suboptimal (Thiede et al., 2003; Van de Pol, De Bruin, Van Loon, & Van Gog, 2019).

According to the cue utilization framework (Koriat, 1997), monitoring accuracy[1] depends on the information or *cues* people use when making judgments (see also Fig. 1). Cues are defined as "bits of information that might potentially be drawn upon or referred to […] to inform a judgment" (Snow, as cited in Cooksey, Freebody, & Wyatt-Smith, 2007, p. 431). As stated in the cue utilization framework, those cues that are most *diagnostic* of students' text comprehension should be *used* (and non-diagnostic cues should be ignored) to arrive at accurate judgments (Koriat, 1997). A cue is diagnostic when it is related to the judged outcome; in this case, students' text comprehension. Yet, students and teachers tend to use cues with low diagnosticity levels (e.g., Van de Pol et al., 2019; Thiede, Griffin, Wiley, & Anderson, 2010). For example, students often use surface cues (e.g., text length) or experiential cues originating from one's subjective experience in performing tasks (e.g., the effort needed to read a text). However, text length or whether a text is read effortlessly does not necessarily predict one's text comprehension and is thus not necessarily diagnostic (Thiede et al., 2010). Thus, using this information does not necessarily help students to arrive at accurate monitoring judgments.

An effective way to improve students' monitoring accuracy is asking students to complete the so-called generative activities such as completing diagrams or concept maps about causal relations in texts (e.g., Thiede et al., 2010; Van Loon et al., 2014). To improve *teachers*' monitoring accuracy, it appears effective to ask them to inspect the results of those completed activities before making judgments about students' understanding (e.g., Van de Pol et al., 2019). Having students perform and teachers inspect the results of generative activities focuses their attention on diagnostic cues. When those diagnostic cues are used to make monitoring judgments, these tend to be more accurate than when these cues are not used (e.g., Van de Pol et al., 2019; Thiede et al., 2010). In this review, we synthesize research that has tested this premise in education with complex text comprehension tasks. We focus on generative activities that require students to transfer text into a visual representation of a that text.

[1] Research on students' text comprehension often refers to "metacomprehension accuracy."

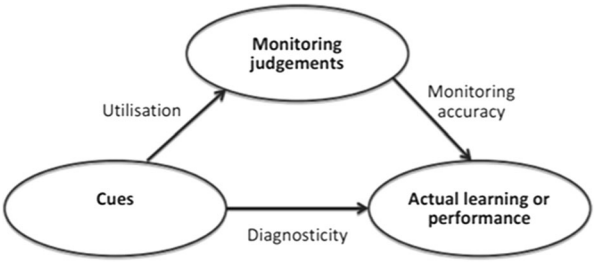

Fig. 1 Relations between cues, monitoring judgments, and actual learning. Reprinted from "Monitoring and regulation of learning in medical education: the need for predictive cues" by B. H. de Bruin, J. Dunlosky, and R. B. Cavalcanti, 2017, *Medical Education, 51*, p. 578. Copyright 2017 by John Wiley & Sons Ltd. and The Association for the Study of Medical Education. Reprinted with permission

Mapping and drawing are appropriate for learning materials that contain several key concepts and (causal or spatial) relations between these concepts. Mapping refers to "a collection of techniques in which a learner converts printed or spoken text into a spatial arrangement of words and links among them" (Fiorella & Mayer, 2016, p. 722). We focus on two mapping techniques: concept mapping and diagramming. Concept mapping refers to creating or completing maps that represent relations between key concepts from a text. Diagramming refers to creating or completing diagrams about causal relations in a text. Drawings are pictorial representations of the text content. These generative activities provide unique opportunities for studying cue diagnosticity and cue utilization, given that the products of these generative activities (e.g., diagram, drawing) can be reliably coded with regard to several cues, for example, the number of correct elements that a product contains (e.g., Van Loon et al., 2014). These coded cues can be related to students' or teachers' judgments to make inferences about presumed cue utilization, and to students' test performance to determine cue diagnosticity (cf. Van Loon et al., 2014). Therefore, we focus on mapping and drawing in this review. Effects of a wider array of generative activities on students' relative monitoring accuracy (excluding cue utilization and cue diagnosticity) can be found in Prinz, Golke, and Wittwer (this special issue).

The first aim of this review is to synthesize research on mapping and drawing that focuses on students' monitoring, regulation, learning, cue utilization, and/or the diagnosticity of cues. The second aim is to discuss whether having access to students' maps and drawings improves *teachers*' monitoring and regulation accuracy, and cue utilization. The third aim is to explore how the effort involved in mapping and drawing affects students' monitoring and regulation. The effort with which texts are read seems to be used as a cue by students (e.g., Schleinschok, Eitel, & Scheiter, 2017), whereby students generally give higher judgments when having completed the learning task effortlessly (Koriat & Ma'ayan, 2005). However, whereas effort invested in reading texts tends to be a low-diagnostic cue that should be avoided (Thiede et al., 2010), the effort invested in generative activities after reading texts, could potentially be a diagnostic cue that could improve students' monitoring accuracy. According to the Effort Monitoring and Regulation (EMR) Framework (De Bruin et al., n.d., this special issue), which synthesizes theory on self-regulation of learning and cognitive load theory, it is important to increase our understanding how effort cues—by itself and in combination with other cues—affect monitoring and regulation of learning. This will help to optimize cue utilization activities and promote monitoring and regulation of text learning. For this third aim, we focus on students, as teachers often do not have insight into students' effort.

Finally, we describe three major future directions in research on mapping and drawing: improving measurements of (a) cue diagnosticity and (b) cue utilization, and (c) discussing how the use of effort as a cue is potentially affected during mapping and drawing and how this cue can be best measured. Future research on these central topics can help in gaining a better understanding of the monitoring and regulation process and how to improve this. Before we address our aims, we first discuss measures of monitoring and regulation accuracy.

Measures of Monitoring and Regulation Accuracy

Monitoring accuracy is mostly expressed in terms of relative or absolute accuracy (cf. Thiede et al., 2019 and Griffin et al., 2019 for extensive discussion). Relative accuracy, often expressed with gamma correlations (G), indicates to what extent one knows which texts are understood better *relative* to other texts (values closer to + 1 indicate higher accuracy). If Julia scores 5 points out of 10 on text A and 7 out of 10 on text B and she thinks she will score 5 points on text A and 6 points on text B, her relative accuracy is perfect because she knows that she understands text B better than text A. However, while her *absolute* accuracy is perfect for text A, it is not for text B. Two common measures of absolute accuracy are *bias* and *absolute deviation.* For bias, the direction of the deviation is taken into account: positive values indicate overestimation and negative values underestimation. For absolute deviation, the direction of the deviation is not taken into account. In our example, the bias score is − 1, (underestimation) and the absolute deviation is 1 (or 9% deviation[2]).

Regulation accuracy is mostly measured in terms of *relative* accuracy (e.g., Schleinschok et al., 2017; Van Loon et al., 2014). Relative regulation accuracy is measured in two ways: monitoring-based regulation (e.g., Van Loon et al., 2014) or comprehension-based regulation (e.g., Schleinschok et al., 2017; Van de Pol et al., 2019). Monitoring-based regulation is determined by relating students' or teachers' monitoring judgments to their restudy decisions. Regulation is considered accurate when those texts are selected for restudy that have received lower monitoring judgments. That is, this measure indicates to what extent students'/teachers' monitoring judgments are related to their regulation decisions, and if so, more accurate monitoring would be expected to be related to more accurate regulation. Comprehension-based regulation is determined by relating students' or teachers' restudy decisions to students' actual comprehension (before actual restudy has taken place). Regulation is considered accurate when those texts are selected for restudy that are understood least well. That is, this measure indicates if regulation decisions are in line with students' performance, regardless of the accuracy of students'/teachers' monitoring judgments of students' performance. Gamma correlations closer to − 1 indicate higher accuracy. Van de Pol, van den Boom-Muilenburg, and van Gog (n.d) determined *absolute* comprehension-based regulation accuracy based on the approach of Baars, Van Gog, de Bruin, and Paas (2014a). Here, students' test scores are compared to students' or teachers' restudy selections. "Low test scores combined with decisions to restudy a text or high test scores combined with decisions to not restudy a text resulted in high regulation accuracy whereas low test scores combined with decisions to not restudy a text or high test scores combined with decisions to restudy a text resulted in low regulation accuracy." Van de Pol, van den Boom-Muilenburg, & van Gog n.d.). If students, for example, can score four points on a test and a student scores 0 points, then a student/teacher

[2] The scale ranges from 0 to 10 (11 scale points), so 1 point deviation is (1/11) × 100 = 9%

would receive one point on the regulation accuracy scale if they indicated that the student would have to restudy the text (because this would be accurate), and zero points if they indicated that the student would not have to restudy the text. If the student scored 3 points, the student/teacher would receive 0.25 points if they did select the text for restudy and 0.75 points if they did not. Values closer to 1 indicate higher accuracy. In this review, we will explicate per study which measure(s) of monitoring and regulation accuracy have been used.

Effects of Generative Activities on Students' Monitoring and Regulation Accuracy, and Learning

Research on self-regulated learning from texts has made use of a "generation paradigm" to improve monitoring accuracy. Such generative activities have been proposed as a means to enhance deeper learning, i.e., leading to better performance (Fiorella & Mayer, 2016). However, in studies on self-regulated learning, these activities are used particularly to improve monitoring and regulation accuracy. Within this paradigm, students first read one or more text(s), engage in generative activities (e.g., make a drawing), make judgments about their test scores, make restudy decisions, and take a test. Often, students do not get the opportunity to restudy texts, because this would confound the relation between monitoring judgments and test performance (monitoring accuracy), unless an extra test is included in the trial (cf. Thiede et al., 2003). In that case, after the first test, students engage in restudying the texts they indicated they should restudy and take a second test. This allows for not only establishing whether more accurate monitoring (i.e., the relation between monitoring judgments and performance on the first test) leads to more accurate regulation but also whether more accurate regulation actually leads to better learning (i.e., performance on the second test; for examples of studies using this two-test paradigm with generative activities, see Kostons & De Koning, 2017; Schleinschok et al., 2017).

The idea of research using generation paradigms is that students self-generate cues that provide them with insight into their text comprehension. Engaging in generative activities promotes monitoring accuracy because activities give students insight into their situation model (e.g., Thiede, Griffin, Wiley, & Redford, 2009). Kintsch (1994) posits that multiple levels of text representations can be constructed during reading: a surface level (representing exact words), a text-base level (representing meaning of sentences), and a situation-model level (representing the gist of a text). Complex texts are only well understood if readers create a high quality situation model representation, by relating different ideas in the text to one another (Griffin et al., 2019; Wiley, Thiede, & Griffin, 2016). Thus, monitoring accuracy for complex texts presumably increases if generative activities provide readers insight into the quality of their situation model.

The generative activities can be performed *during* reading (concurrent approach; e.g., Thiede et al., 2019; Wiley, 2019), *immediately* after having read each text (immediate approach; e.g., Van Loon et al., 2014), or at a delay, *after* having read *all* texts (delayed approach; e.g., Schleinschok et al., 2017), see Fig. 2. Early studies using generative activities (e.g., Thiede & Anderson, 2003; Thiede et al., 2003) found that generative activities only improved relative monitoring accuracy when performed at a delay (as opposed to immediately). One explanation for why delayed generative activities might lead to higher monitoring accuracy than immediate activities is that when performed immediately after reading, information about the text content may still come to mind easily; memory decay has not taken place yet. This is detrimental for monitoring accuracy

because students still have much information from the text accessible in their working memory immediately after reading (and might therefore estimate their test performance to be high), but some of these memory traces will have decayed by the time the test is taken. In contrast, when performed at a delay (in which other information is processed), the generation task helps people to reflect on how successful one is at retrieving situation model information from long-term memory (Griffin et al., 2019; Thiede et al., 2019). As this also resembles the situation during the comprehension test (where information has to be retrieved from long-term memory), the cues students gain from engaging in delayed generative activities will be more predictive of test performance (Griffin et al., 2019; Thiede et al., 2019).

More recently, some studies have used a concurrent approach (e.g., Redford, Thiede, Wiley, & Griffin, 2012; Thiede et al., 2019; Fig. 2). In this approach, students perform generative activities during reading and students are provided with "instructions for reading texts that promote construction of the situation model—connecting ideas in a text to one another and to one's prior knowledge." (Thiede et al., 2019, p. 3). This is hypothesized to promote (relative and absolute) monitoring accuracy because the construction of a high quality situation model makes cues that are related to this situation model more salient. And because cues that are related to a student's situation model are more diagnostic, they help students arriving at more accurate judgments (cf. Thiede et al., 2019).

Although early research shows benefits of generative activities for improving monitoring and regulation accuracy (e.g., Thiede et al., 2003), most studies did not explicitly assess which cues became accessible to students from engaging in generative activities, how these cues were related to their later test performance (i.e., cue diagnosticity), and whether these cues were related to their judgments (i.e., cue utilization). Exceptions are studies by Anderson and Thiede (2008) and Thiede and Anderson (2003), in which cues yielded by summarizing were coded and cue diagnosticity and cue utilization were measured. Lately, these questions were investigated in more detail using mapping and drawing (e.g., Schleinschok et al., 2017; Van Loon et al., 2014).

Effects of Mapping and Drawing In the last decade, researchers have used the generation paradigm for mapping and drawing. For this review, we discuss studies that (a) addressed students' and/or teachers' monitoring/regulation accuracy, (b) used mapping or drawing as a generative activity, and (c) reliably coded one or more cues from the maps or drawings to infer cue diagnosticity and/or cue utilization. Studies that meet these criteria are listed in Table 1.

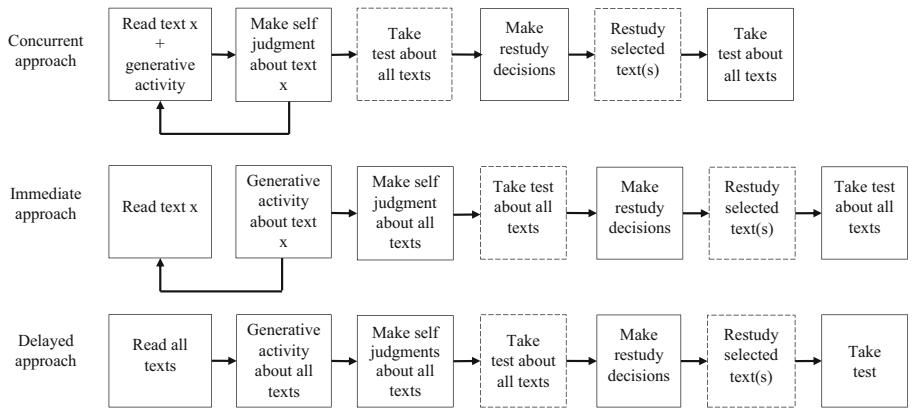

Fig. 2 Generation paradigm for students. Activities in boxes with dashed lines are often omitted

Table 1 Studies on mapping and drawing included in this review

Study	Participants	Generative activity	Experimental conditions	Main findings regarding monitoring accuracy
Thiede et al., 2010	University students	Concept mapping while reading five texts	No activity + immediate JOL, no activity + delayed JOL, concept mapping + JOL	Relative accuracy: concept mapping > immediate and delayed
Redford et al., 2012	7th grade	Concept mapping while reading three texts	Exp1: concept-mapping/rereading Exp2: concept map construction, concept map provision, control	Experiment 1 (relative accuracy): concept-mapping = rereading Experiment 2 (relative accuracy): concept-mapping > control; concept mapping = concept map provision; concept map provision = control
Van Loon et al., 2014	9th grade	Diagramming	Immediate diagramming, delayed diagramming, no diagramming	Relative accuracy: delayed diagramming > no diagramming; immediate diagramming = delayed diagramming; immediate diagram = no diagramming
Kostons & De Koning, 2017	4th and 5th grade	Visualization	Long drawing condition, brief drawing condition, control	Absolute accuracy: long-drawing > brief drawing; long--drawing = control; brief-drawing = control
Schleinschok et al., 2017	University students	Drawing after reading each text paragraph	Drawing, control	Absolute accuracy: drawing = control Relative accuracy, gamma correlations: drawing = control Relative accuracy, MLA: drawing > control
Van de Pol et al., 2019	Students & teachers	Completing and drawing diagrams	Students: delayed diagram completion, delayed diagram drawing, control Teachers: name-only, name+completed diagram, name+drawn diagram.	Students (relative accuracy): Delayed diagramming/drawing > control; delayed diagramming = delayed drawing Teachers (relative accuracy): name-only = name+completed diagram = name+drawn diagram.
Wiley 2019	Undergraduates	Drawing while reading a text	Drawing (sketching), photograph, no-image, diagram provision, note-taking, animation	Absolute accuracy (deviation): drawing > photograph, no-image, animation condition Absolute accuracy (bias): drawing > photograph/no-image
Thiede et al., 2019	5th grade	Drawing while reading texts	Organizational, representational	Relative accuracy: organizational > representational
Authors (SubmittedA)	Secondary education teachers	Students' completed diagrams	Name-only, name+diagram, diagram-only	Absolute accuracy (deviation): name+diagram > diagram-only; diagram-only = name-only; name-only = name+diagram
Authors (SubmittedB)	Secondary education teachers	Students' completed diagrams	Name+diagram	• Using non-diagnostic student cues (general reading comprehension level, grades other subjects, nationality, extraversion, and IQ) resulted in more overestimation. • Higher accuracy (absolute deviation and bias) when estimations of correct relations in students' diagrams and students' general effort in class were more accurate

Diagramming Van Loon et al. (2014) conducted the first study using diagramming as a generative activity to improve students' relative monitoring accuracy. They asked 9th grade students to read six texts containing cause-and-effect relations. A diagram completion task was designed to focus students' attention on their understanding of these relations. For this task, diagrams consisting of empty boxes connected by arrows had to be completed. Three conditions were compared; one condition completed diagrams immediately after reading a text, one condition completed them at a delay, and a control condition did not complete diagrams. Then, participants made monitoring judgments, restudy selections, and took a causal relations test. In the delayed-diagramming condition, students were better able to discriminate between well and less well learned texts ($G = 0.56$) than in the control condition ($G = 0.07$). Other comparisons were not significant. Moreover, monitoring-based (relative) regulation accuracy was high: participants most often selected texts for restudy that were judged as less well-learned ($G_{overall\ mean} = -0.65$). Furthermore, students in the immediate-diagramming condition had higher test scores compared to the other conditions. Thus, delayed diagramming affected monitoring accuracy, not performance. As monitoring was most accurate in the delayed-diagramming condition and as there was a strong relation between monitoring and regulation decisions, students in the delayed-diagramming condition may ultimately benefit most when given a restudy opportunity.

The design used in Van Loon et al. (2014) enabled measurement of cues which became accessible through diagram completion (also see Thiede et al., 2010). By coding the diagrams with respect to, for instance, the number of empty boxes (i.e., omissions), Van Loon et al. (2014) analyzed the cue diagnosticity and cue utilization. Table 2 shows the cue diagnosticity values (relation between cues and test performance) and students' cue utilization (relation between cues and judgments) for all cues measured. For both diagram conditions, the number of correct relations and the number of omissions were diagnostic for test performance. Interestingly, in the delayed diagram condition, there was a strong correlation between these diagnostic cues and students' judgments of learning, suggesting that students in this condition may have used these diagnostic cues when making their judgments. These analyses thus seemed to confirm theoretical assumptions that delayed generation explicitly focuses students' attention on diagnostic cues.

A study by Van de Pol et al. (2019) replicated the findings on the benefits of delayed diagram completion on relative monitoring accuracy compared to a no-diagram condition. Furthermore, they investigated effects of a delayed diagram *drawing* instruction, where 9th grade students also drew the structure of the diagram, rather than only complete a pre-structured diagram. The delayed diagram completion condition ($G = 0.43$) and the delayed diagram drawing condition ($G = 0.28$) showed higher relative monitoring accuracy than the no-diagram condition ($G < 0.01$). The difference between the diagram conditions was not significant. Although the number of correct relations was diagnostic in both diagramming conditions, this cue related more strongly to students' judgments in the diagram-completion condition than in the diagram-drawing condition, suggesting that students in the diagram-completion condition may have made more use of this diagnostic cue than in the diagram-drawing condition. Thus, the diagram-drawing condition could not benefit as much from diagnostic diagram cues as the diagram-completion condition. Furthermore, and similar to Van Loon et al. (2014), monitoring-based (relative) regulation accuracy was high, with no differences between conditions ($G = -0.73$): texts that were judged as less well learned were more often selected for restudy. Additionally, comprehension-based (relative) regulation accuracy was moderate ($G_{overall} = 0.3$), with no differences between conditions. So although students'

Table 2 Cues based on the generative activities measured in the reviewed articles, their diagnosticity and cue utilization

Cues	Explanation of the cue	Cue diagnosticity[a]	Cue utilization	
			Students	Teachers
Diagramming				
Number of correct relations	Number of correct combinations of two diagram boxes	$r_{\text{delayed completion}} = 0.49^{*c}/0.39^{*d}$ $r_{\text{immediate completion}} = 0.53^{*c}$ $r_{\text{delayed drawing}} = 0.36^{*d}$	$G_{\text{delayed completion}} = 0.59^{*c}/0.57^{*d}$ $G_{\text{immediate completion}} = 0.23^{*c}$ $G_{\text{delayed drawing}} = 0.26^{*d}$	$G_{\text{Name + completed diagram}} = 0.54^{*d}/r = 0.59^{*b}$ $G_{\text{Name + drawn diagram}} = 0.39^{*d}$
Number of completed boxes	Number of completed diagram boxes	$r_{\text{delayed completion}} = 0.26^{*d}$ $r_{\text{delayed drawing}} = 0.28^{*d}$	$G_{\text{delayed completion}} = 0.71^{*d}$ $G_{\text{delayed drawing}} = 0.44^{*d}$	$G_{\text{Name + completed diagram}} = 0.68^{*d}$ $G_{\text{Name + drawn diagram}} = 0.57^{*d}$
Omissions	Relevant information from the text not present in the diagram	$r_{\text{delayed completion}} = -0.39^{*c}/-0.26^{*d}$ $r_{\text{immediate completion}} = -0.24^{*c}$ $r_{\text{delayed drawing}} = -0.28^{*d}$	$G_{\text{delayed completion}} = -0.64^{*c}/-0.71^{*d}$ $G_{\text{immediate completion}} = -0.50^{*c}$ $G_{\text{delayed drawing}} = -0.45^{*d}$	$G_{\text{Name + completed diagram}} = -0.67^{*d}/\ r = -45^{*b}$ $G_{\text{Name + drawn diagram}} = -0.56^{*d}$
Commission errors	Information in the diagram not presented in the text	$r_{\text{delayed completion}} = -0.17^{*c}/-0.14^{*d}$ $r_{\text{immediate completion}} = -0.30^{*c}$ $r_{\text{delayed drawing}} = -0.05^{d}$	$G_{\text{delayed completion}} = -0.16^{c}/0.10^{d}$ $G_{\text{immediate completion}} = 0.00^{c}$ $G_{\text{delayed drawing}} = 0.12^{d}$	$G_{\text{Name + completed diagram}} = -0.04^{d}/r = 0.29^{*b}$ $G_{\text{Name + drawn diagram}} = 0.35^{*d}$
Factual information	Factual information presented in the text	$r_{\text{delayed completion}} = -0.09^{c}$ $r_{\text{immediate completion}} = -0.22^{*c}$	$G_{\text{delayed completion}} = 0.20^{c}$ $G_{\text{immediate completion}} = -0.11^{c}$	$r_{\text{name + completed diagram}} = 0.08^{*b}$
Extensiveness formulations	Average no. of words over all diagram boxes			$r_{\text{name + completed diagram}} = 0.38^{*b}$
Completion time	Time needed to complete diagram			$r_{\text{name + completed diagram}} = 0.03^{b}$
Drawing				
Correct idea units	Correctly represented major/minor idea units	$r = 0.65^{*e\text{-}exp1}/r = 0.78^{e\text{-}exp2}$	$G = 0.44^{*e\text{-}exp1}/G = 0.18^{*e\text{-}exp2}$	
Number of elements	Number of elements from the text presented in the drawing	$r = 0.25^{*f}$; $G = 0.20^{g}$	$G = -0.11^{g}$	
Number of big ideas	Big ideas in the text	$G_{\text{representational}} = -0.14^{g}$ $G_{\text{organizational}} = 0.42^{*g}$	$G_{\text{representational}} = 0.10^{g}$ $G_{\text{organizational}} = 0.22^{*g}$	
Detail level	Number of details present in the drawing	$r = -0.27^{*f}$		
Number of actions	Actions described in the text	$G_{\text{representational}} = -0.15^{g}$ $G_{\text{organizational}} = 0.13^{g}$	$G_{\text{representational}} = -0.16^{g}$ $G_{\text{organizational}} = 0.24^{*g}$	
Number of related elements	Elements presented in the text	$G_{\text{representational}} = -0.01^{g}$ $G_{\text{organizational}} = 0.18^{g}$	$G_{\text{representational}} = 0.14^{g}$ $G_{\text{organizational}} = 0.07^{g}$	
Number of novel elements	Elements not presented in the text but related to the topic	$G_{\text{representational}} = -0.12^{g}$ $G_{\text{organizational}} = 0.13^{g}$	$G_{\text{representational}} = 0.40^{g}$ $G_{\text{organizational}} = 0.10^{g}$	
Number of unrelated elements	Elements not presented in the text and not related to the topic	$G_{\text{representational}} = -0.15^{g}$ $G_{\text{organizational}} = -0.01^{g}$	$G_{\text{representational}} = 0.07^{g}$ $G_{\text{organizational}} = -0.11^{g}$	
Semantic relations	How text and drawing are related (vaguely/representational/organizational/interpretational)	$G_{\text{representational}} = 0.03^{g}$ $G_{\text{organizational}} = 0.22^{*g}$	$G_{\text{representational}} = 0.16^{g}$ $G_{\text{organizational}} = 0.24^{*g}$	

Table 2 (continued)

Cues	Explanation of the cue	Cue diagnosticity[a]	Cue utilization	
			Students	Teachers
Connections	Drawing includes information from reader's background/prior knowledge	$G_{representational} = -0.09$[g] $G_{organizational} = 0.66$[g]	$G_{representational} = 0.73$*[g] $G_{organizational} = 0.60$*[g]	
Captions and labels	"parts of a diagram, the steps in a process or both0." (Thiede et al., 2019, p. 8)	$G_{representational} = -0.02$[g] $G_{organizational} = 0.15$[g]	$G_{representational} = -0.02$[g] $G_{organizational} = 0.07$[g]	
Systematicity	How well the situation model of the system in the text is represented in the drawing	$G_{representational} = 0.22$[g] $G_{organizational} = 0.24$*[g]	$G_{representational} = 0.21$[g] $G_{organizational} = 0.27$*[g]	
Concept mapping				
Number of nodes	Elements of a concept map	$r_{limited\ instructions} = NS$[h] $r_{extensive\ instructions} = 0.44$[h-exp2]		
Paragraph structure/use	Number of nodes reflecting the paragraph structure	$r_{limited\ instructions} = NS$[h] $r_{extensive\ instructions} = 0.27$*[h-exp2]		
Redundant nodes	Nodes not needed to understand the text	$r_{limited\ instructions} = -0.35$*[h-exp1] $r_{extensive\ instructions} = NS$[h]		
Number of connections	Connections made between nodes	$G = 0.38$*[i]	$G = 0.32$*[i]	
Multiple link nodes	Number of nodes having multiple links	$r_{limited\ instructions} = NS$[h] $r_{extensive\ instructions} = NS$[h]		

r Pearson correlation, *G* intra-individual gamma correlation, *NS* not significant

* Correlation is significantly different from zero ($p < 0.05$)

[a] Values closer to + 1/− 1 indicate high diagnosticity levels; closer to 0 indicate low diagnosticity levels

[b] Authors (SubmittedB)

[c] Van Loon et al. (2014)

[d] Van de Pol et al. (2019)

[e] Schleinschok et al. (2017)

[f] Kostons and De Koning (2017)

[g] Thiede et al. (2019)

[h] Redford et al. (2012): this study only reported significant correlations

[i] Thiede et al. (2010)

monitoring was more accurate in the diagramming conditions, comprehension-based regulation accuracy was not necessarily higher in these conditions. This may hamper students' learning when given the opportunity to restudy, because their restudy decisions are not fully in line with their actual comprehension. So unless monitoring-based regulation accuracy is perfect (G = 1), it is informative to calculate comprehension-based regulation accuracy as this can explain or predict effects of restudy on students' learning. Finally, there were no differences in test scores between the conditions (which was also not necessarily expected, as diagramming was mostly expected to affect monitoring and regulation).

These studies show the merit of measuring both cue diagnosticity and cue utilization to help understand the effects of different instructions. Delayed diagram completion showed the greatest improvement in students' monitoring accuracy. And although several diagram cues seemed diagnostic in all diagramming/drawing conditions, students in the delayed-diagram completion condition may have used these most often (based on the correlation between judgments and cues). Completing diagrams at a delay thus seems to help students focus on and possibly use diagnostic cues that are indicative of the quality of their situation model. Yet, it is not clear why more accurate monitoring is not necessarily accompanied by more accurate comprehension-based regulation.

Concept Mapping Two studies have investigated the effect of concept mapping on students' monitoring accuracy. Thiede et al. (2010) investigated to what extent university students' relative monitoring accuracy benefitted from engaging in concept mapping while reading texts. At-risk readers participated in each condition: (1) making immediate judgments after having read each of five texts and then taking a test; (2) making delayed judgments by first reading five texts, then making five judgments, and then taking a test; and (3) constructing concept maps while reading each of five texts, making a judgment about that text without access to their concept map, and taking a test for that text. For this latter condition, students received 8 × 50-min training on constructing concept maps. Students' relative monitoring accuracy was higher in the concept map condition (G ± 0.68[3]) than in the immediate (G ± 0.30) and delayed (G ± 0.33) conditions. Additionally, students' test performance was higher in the concept map condition than in the other conditions. The number of appropriate connections in the concept map were diagnostic of (correlated with) students' performance and students probably used this cue in making their judgments (indicated by the correlation between cues and students' judgments; see Table 2), which may explain the benefits of concept mapping.

Redford et al. (2012) investigated whether students' relative monitoring benefitted from concept mapping while having access to their concept maps during monitoring. In experiment 1, 7th grade students either read three texts, reread the texts, and made judgments (reread condition), or read three texts while making concept maps, made judgments while having access to their concept maps, and took a test (concept-mapping condition). Students received 45–60-min concept mapping training. Monitoring accuracy was not higher in the concept mapping condition (G ± 0.05) than in the reread condition (G ± − 0.4). None of the cues measured (e.g., number of nodes and paragraph structure) was diagnostic (Table 2). Cue utilization was not measured. In experiment 2, they implemented more elaborate concept mapping instructions (120 min) focusing on how to use concept maps for improving text understanding rather than summarizing the text. Students in the concept-mapping condition

[3] We used the ± sign when the exact number was not stated in the article (e.g., results are provided in a figure).

had higher monitoring accuracy (G ± 0.34) than students who did no extra activity (G ± − 0.24). Monitoring accuracy of students who were provided with completed concept maps (G ± 0.17) was not different compared to the other two conditions. Paragraph-use in the concept maps was diagnostic for students' test performance (Table 2). Test performance did not differ between conditions.

In sum, these studies show that students' monitoring accuracy can benefit from concept mapping. Extensive instructions focusing on representing relations in the concept maps may be important to benefit from concept mapping, especially for young learners. Additionally, it seems beneficial to ask students to make judgments without access to their concept maps, possibly because this resembles the test situation more in which students do not have access to their concept maps either and also have to retrieve information actively.

Drawing Other studies have focused on drawing as a generative activity. In two studies, students were provided with specific and elaborate drawing instructions to focus their attention on diagnostic cues, namely, the relations in the text. In Kostons and De Koning (2017), two drawing conditions (short drawing during reading texts and long drawing during and after reading texts including taking the test) were compared to a no-drawing condition. Students' (4th/5th grade) absolute monitoring accuracy was higher in the long-drawing condition (25.2% deviation) than in the brief-drawing condition (34.2% deviation), but not more than in the no-drawing condition (31.4%). Cue utilization was not measured. The level of detail and the number of elements in the drawings were diagnostic (Table 2) and drawings in the long-drawing condition had more elements and details than in the brief-drawing condition. This may explain differences in monitoring accuracy. Overall monitoring-based (relative) regulation accuracy was moderate ($r = -0.36$). However, comprehension-based (relative) regulation accuracy was low ($r = 0.12$). Thus, although students' regulation decisions were to some extent aligned with their judged understanding, the restudy decisions were not aligned with their actual understanding. It is therefore not surprising that no effects of actual restudy on students' learning were found; students' test scores after restudy (and before restudy) did not differ between conditions.

Thiede, Wright, Hagenah, and Wenner (2019b) asked 5th grade students to make drawings during reading. One condition was explicitly instructed to include information about the relations between text ideas (organizational drawing condition), whereas the other condition was only told to represent the text with their drawing (representational drawing condition). The organizational drawing condition showed higher relative monitoring accuracy (G = 0.51) than the representational drawing condition (G = − 0.03). Students' test scores are not reported in this study. Although students in both conditions seem to have used relational information (indicated by relations between cues and students' judgments), this cue was only diagnostic of performance for the organizational drawing condition (Table 2). Thus, only students in the organizational condition could actually benefit from the drawing instruction. This way, students are less likely to mainly attend to and reproduce surface and text base details during drawing, and instead, to focus on the gist of the text by trying to relate elements from the text. This focus on the gist of the text gives them a good idea of the quality of their situation model which helps them in monitoring their understanding accurately.

Two other studies gave more general drawing instructions, that is, to draw the content of the text. Wiley (2019) investigated the effect of drawing while reading a text, compared to five other conditions on students' monitoring accuracy. In the other conditions, undergraduate students had to generate written information (note-taking), were provided with a visualization

(static or animated diagram, photograph) or were provided with no visualization (no-image). Students' absolute monitoring accuracy was higher in the drawing condition (± 5% deviation) than the photograph, no-image, and animation conditions (± 15–25% deviation). Additionally, students in the drawing condition were less overconfident than in the photograph and no-image conditions. Students in the drawing condition had higher test scores than students in the photograph and no-image conditions. Two cues were coded: interest in the text topic and perceived helpfulness of the text to understand the topic. Cue diagnosticity was not measured, but based on other studies it can be assumed that these cues are not diagnostic (e.g., Van de Pol et al. (n.d.). Students seem to have based their monitoring judgments on their interest. However, we do not know whether the improved monitoring accuracy of the drawing condition is related to better cue utilization (e.g., less use of interest), given that no (correlational) cue utilization scores were presented per condition.

Schleinschok et al. (2017) asked university students to make a drawing of a text *after* reading a paragraph, without seeing the text. In the experimental condition, students were asked to draw the content of each of five paragraphs (general drawing instructions), make a judgment after finishing the drawing, make restudy decisions, (+actual restudy of the selected paragraphs in experiment 2), and take a test after having finished all drawings and judgments. Students in the control condition did not draw. Students' judgments were close to their actual test scores (6% and 8% deviation in the control and drawing condition respectively), but there were no differences between conditions. Relative monitoring accuracy also did not differ between conditions and their ability to discriminate between paragraphs was only moderate ($G_{\text{drawing condition}} = 0.29$, $G_{\text{control condition}} = 0.27$). Follow-up analyses showed that judgments may have been more predictive of performance in the drawing condition than in the control condition. Additionally, monitoring-based (relative) regulation accuracy was high ($G_{\text{overall mean}} = -0.60$,) and comprehension-based (relative) regulation accuracy was moderate ($G_{\text{overall mean}} = -0.26$). However, there were no differences in regulation accuracy between conditions. Further, no difference was found between conditions in performance before and after actual restudy. A lack of difference after actual restudy may be due to the fact that students made little use of the restudy opportunity. Additionally, although there was an indication that relative monitoring was more accurate in the drawing condition, comprehension-based regulation accuracy was not, so students did not necessarily select those paragraphs for restudy that they understood less well. The number of correct elements in the drawings appeared diagnostic of students' performance and it seemed that participants have used this cue in making monitoring judgments (Table 2). However, given that cue utilization scores were not provided per condition, we cannot determine whether cue utilization helps to understand differences between conditions.

In sum, drawing seemed to foster monitoring accuracy mainly when students received specific and elaborate instructions to focus on the relations in the text, at least when looking at relative accuracy. For absolute accuracy, findings are mixed. Although given specific instruction, drawing did not improve absolute monitoring accuracy in Kostons and De Koning (2017) while general instructions did in Wiley (2019), but did not in Schleinschok et al. (2017). The difference in findings between the latter two studies may be related to the fact that in Schleinschok et al. (2017), students drew about each paragraph whereas in Wiley (2019), students drew about the whole text. Probably, the quality of the situation model benefits most in the latter case, because then, the drawings also include relations between paragraphs. Only Thiede et al. (2019) provided full information about students' cue utilization

and cue diagnosticity per condition and showed that cue diagnosticity differed between instructional conditions.

Conclusion Regarding relative accuracy, concept mapping during reading with no access to the concept map during monitoring was most effective in promoting students' monitoring accuracy (Thiede et al., 2010). Diagram completion (Van Loon et al., 2014; Van de Pol et al., 2019) and organizational drawing (Thiede et al., 2019) were also effective. Yet, we cannot conclude that mapping or drawing is *always* effective; there seem to be boundary conditions.

First, the explicitness and extensiveness of the instructions seem to matter for concept mapping and drawing to be effective for monitoring accuracy. Only when students are instructed to focus on relations in the text in their concept maps or drawings, do the students' products seem to contain diagnostic cues which are needed for accurate judgments. For diagramming, such extensive instructions are not needed because the diagram that students complete is pre-structured and focuses students' attention on the relations in the text.

Second, several studies (Thiede et al., 2019; Van de Pol et al., 2019; Van Loon et al., 2014) have shown that only generating diagnostic cues is not sufficient. These cues also need to be *used* to become more accurate. When developing interventions, attention should thus be paid to promoting the generation of diagnostic cues and students' actual *use* of these cues.

Third, studies with a delayed approach (e.g., Van Loon et al., 2014) and a concurrent approach (e.g., Thiede et al., 2019) have shown improvements in students' monitoring accuracy. Yet, only when students did *not* have access to the products of their generative activities when making judgments, did their monitoring accuracy benefit. Thus, some form of retrieval, namely at least when making the judgments, might be necessary for promoting students' monitoring accuracy.

Although monitoring accuracy seems to benefit from generative activities, students' actual learning after restudy does not. Some studies have shown that although monitoring becomes more accurate, comprehension-based regulation accuracy is low (e.g., Van de Pol et al., 2019). This means that although students' regulation decisions may be related to their monitoring judgments (i.e., comprehension-based regulation accuracy), these regulation decisions do not match their comprehension so students may still study texts that they already understand or not study texts that they do not understand.

Finally, of the studies that did not include an actual restudy phase, two studies found that students who engaged in the generative activity not only had higher monitoring accuracy, they also scored higher on the test than students who did not (Thiede et al., 2010; Wiley, 2019). In the study of Wiley (2019), we cannot be sure that the higher absolute monitoring accuracy does not stem from the fact that students' understanding also increased by engaging in the generative activity. That is, students generally overestimate their performance (e.g., Foster, Was, Dunlosky, & Isaacson, 2017). If students perform better, there is less room for overestimation and this may be a reason why students' absolute monitoring accuracy becomes more accurate. Yet, other studies (Van de Pol et al., 2019; Van Loon et al., 2014) did not find an effect of the generative activity on learning, but did find an effect on students' monitoring accuracy. Thus, engaging in generative activities (in this case diagramming) can specifically foster monitoring accuracy. Prinz et al. (this special issue), also found in their meta-analysis that the effect of engaging in generative activities on students' relative monitoring accuracy is stronger than the effect on students' learning. However, the design of the study and the timing of the generative activity may also play a role. That is, in the study of Wiley (2019), students made drawings during reading. Such concurrent generative activity can deepen the processing

of the text and therefore contribute to better performance. In contrast, delayed generative activities (e.g., Van de Pol et al., 2019; Van Loon et al., 2014) focus on students' retrieval from text and probably do not involve deeper processing or understanding of the text and which may not result in higher direct test performance.

Effects of Viewing Students' Maps and Drawings on Teachers' Cue Utilization, Monitoring, and Regulation Accuracy

Students' maps or drawings generate diagnostic cues not only for students but also for teachers. Although experiential cues about how maps or drawings were completed (e.g., with how much effort) do not become available for teachers upon inspection, diagnostic cues related to the quality of the products do become available. Thus, having access to maps or drawings when judging students' text comprehension may also help teachers in monitoring and regulation of students' learning. This may be useful as teachers determine what their students know by looking at their work (in this case, their maps or drawings; see Van de Pol et al., 2019). This can be done either during lessons in which students work on maps or drawings (e.g., by walking around in the classroom) or in between lessons when assessing students' work (e.g., Smit, Van Eerde, & Bakker, 2013). Based on this, teachers can adapt their instruction or explanation while teaching or their lesson plans for future lessons accordingly, therewith promoting students' learning (Van de Pol et al., 2010; Wittwer & Renkl, 2008).

The generation paradigm for teachers is depicted in Fig. 3. Currently, this paradigm has only been used in the context of diagramming, but it can also be used in future research using other generative activities. In this paradigm, teachers first read all texts. Then, they view a student's generated product (e.g., diagram) about a particular text and, while having access to that product, judge the student's comprehension of that text. This is repeated for each text. Finally, the teacher indicates which text(s) should be restudied before taking a test. Note that in this paradigm, teachers view a student's generated product during monitoring whereas students mostly do not see their product when monitoring their own understanding (see for an exception Redford et al., 2012, using the concurrent approach). For students, making judgments without seeing their maps or drawings makes sense because this resembles the actual test situation in which they also do not have access to their maps or drawings. Yet, for teachers, the situation is different because they judge the text understanding of their students and do not take a test.

This generation paradigm for teachers was developed by and first used in Van de Pol et al. (2019). In this study, secondary education teachers judged the text comprehension of their students while (a) knowing students' names only (making student cues available), (b) knowing students' names and having access to students' completed diagrams, and (c) knowing students' names and having access to students' drawn diagrams (in (b) and (c) student cues and performance cues are made available). Teachers' relative monitoring accuracy did not differ between conditions and was low to moderate ($G_{\text{control condition}} = 0.03$ and $G_{\text{diagramming conditions}} = 0.26$). Teachers' monitoring-based (relative) regulation accuracy was high; teachers' monitoring judgments were highly related to their regulation decisions. Yet, there were no differences between conditions ($G_{\text{overall}} = -0.82$). Comprehension-based (relative) regulation was more

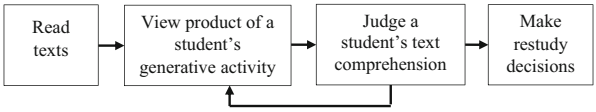

Fig. 3 Generation paradigm for teachers

accurate in the diagramming conditions. That is, teachers more often selected those texts for restudy that students indeed understood less well compared to the control condition ($G_{name+completed\ diagram} = -0.38$; $G_{name+drawn\ diagram} = -0.21$; $G_{control} = 0.00$). Thus, the accuracy of teachers' judgments of students' understanding was not affected by condition, whereas the accuracy of their regulation decisions was. This suggests that monitoring and regulation judgments may be, to some extent, based on different cues.

Teachers in both diagramming conditions seemed to have used the number of completed boxes, correct relations, and omissions in students' diagrams as a cue for their monitoring judgments to a similar extent (based on the correlations between their judgments and cues; see Table 2), which may explain the fact that there was no difference between the diagramming conditions in terms of monitoring accuracy. The finding that teachers' monitoring was not more accurate in the diagramming condition than in the control condition, may be due to the fact that, in all conditions, teachers knew about which student they were making the judgments so general information about the student (e.g., IQ, effort in class, conscientiousness), which is not necessarily diagnostic, was available. This may have hampered teachers' monitoring accuracy in the diagramming condition.

To assess this explanation, Authors (SubmittedA) (n.d.) conducted a follow-up study. Secondary education teachers made judgments about their students under each of the following three conditions: only knowing the student's name (name-only), only seeing anonymized diagrams (diagram-only), or knowing students' names and seeing their diagrams (name+diagram). Teachers thought out loud while making judgments and these think aloud protocols were analyzed with regard to teachers' cue utilization to be able to measure the use of not only diagram cues, but all sorts of performance, student-related, and task cues. Teachers' absolute monitoring accuracy was higher in the name+diagram condition (22.6% deviation) than in the diagram-only condition (28% deviation). In contrast, teachers' regulation was more accurate in the diagram-only condition than in the name-only condition (relative comprehension-based regulation accuracy: $M_{name\text{-}only} = -0.10$, $M_{name+diagram} = -0.31$, $M_{diagram\text{-}only} = -0.29$; absolute monitoring-based regulation accuracy: $M_{name\text{-}only} = 0.43$, $M_{name+diagram} = 0.38$, $M_{diagram\text{-}only} = 0.35$[4]). Exploratory analyses of the think aloud data suggested that teachers regularly misinterpreted diagnostic diagram cues, especially the number of correct relations in the diagrams. Accurately interpreting the cue *manifestation* (e.g., the *actual* number of correct relations) may thus affect monitoring accuracy. With the cue manifestation, we mean the actual value of a cue for a particular student, e.g., a student's actual interest in a text topic or the actual number of commission errors (i.e., the provision of a wrong answer).

Therefore, in a subsequent study, the effect that the ability of judging cue values has on teachers' absolute monitoring accuracy was studied. In Authors (SubmittedB) (n.d.), secondary education teachers judged the text comprehension of their students. While making those judgments, they had information available from which they could deduce cues: students' completed diagrams (giving access to performance cues), students' names (giving access to student cues), and the texts (giving access to task cues). For each judgment, they indicated on a list of 28 cues what cue(s) they used. The list was based on think-aloud data of previous studies (Oudman, Van de Pol, Bakker, Moerbeek, & Van Gog, 2018; Van de Pol et al., 2019) and was complemented with cues from the literature (e.g., Cooksey et al., 2007; Dinsmore & Parkinson, 2013; Mizala, Martínez, & Martínez, 2015). Examples of cues are: correct relations

[4] Closer to 0 is more accurate.

or omissions in the diagram (i.e., performance cues), IQ or gender (i.e., student cues), and text length (i.e., task cues). Teachers were additionally asked to judge the manifestations of the cues they had used. If a teacher for example indicated that they used the cue "IQ," they were subsequently asked to judge this student's score on an IQ test. All cues on the cue list were measured to be able to determine the cue utilization and cue diagnosticity. Mere *use* of diagnostic cues did not promote teachers' absolute monitoring accuracy. That is, when teachers used highly diagnostic cues, their judgments were not more accurate. Rather, *accurately judging cue manifestations* of highly diagnostic cues (e.g., correct relations in the diagram) improved teachers' monitoring accuracy. For example, even when they would know that correct relations are diagnostic and use this cue, this would not help them make more accurate judgments when they cannot judge accurately how many correct relations are actually present in the diagram. Additionally, the use of non-diagnostic student cues (i.e., students' general reading comprehension levels, grades for other subjects, nationality, extraversion, IQ) was found to *hamper* teachers' monitoring accuracy. Determining teachers' (or students') judgment accuracy of cue manifestations may thus be important when investigating and improving monitoring accuracy.

Studies using the generation paradigm for teachers have yielded new insights in teachers' monitoring and regulation of students' comprehension. First, the availability of cues appears to affect monitoring and regulation differently. Teachers' regulation accuracy mainly benefitted from being provided with information from which they could deduce diagnostic cues, but interestingly, not their monitoring accuracy. Second, just providing teachers with information that contains diagnostic cues does not seem enough to promote their monitoring accuracy; they also have to ignore non-diagnostic cues and be able to accurately judge the manifestations of diagnostic cues to become more accurate at monitoring of students' learning.

Effort as a Cue

Our third aim is to discuss how mental effort invested in mapping and drawing is potentially used as a cue for monitoring and regulation, under what circumstances effort may be diagnostic, and how the use of effort as a cue affects students' monitoring accuracy. We will first discuss the use of effort as a cue more generally, before discussing findings on effort and monitoring and regulation accuracy from two studies on drawing.

Research on self-monitoring and self-regulation with problem-solving tasks suggests students use mental effort invested in a task as a cue for monitoring judgments (based on the correlation between judgments and cues): The higher the effort students had to invest, the lower they expected to perform on the test (e.g., Baars, Visser, Van Gog, De Bruin, & Paas, 2013; Baars, Van Gog, De Bruin, & Paas, 2018; Baars, Vink, van Gog, de Bruin, and Paas, 2014b; see also Baars, Wijnia, De Bruin, and Paas, n.d., this special issue). Whereas the mental effort invested in the task might not always be a good cue for judging comprehension, the amount of effort students would need to invest to complete generative activities, might be a good indicator of the quality of their mental model of the solution procedure or –in text learning – their situation model. That is, the amount of mental effort a task requires is associated with students' skill level (i.e., less-skilled students attain lower accuracy in more time and with more effort than more-skilled students who attain higher accuracy in less time and with less effort; see Kalyuga, 2006; Van Gog & Paas, 2008). As such, the amount of effort invested in the generative activity might be a diagnostic cue that could enhance monitoring

accuracy. We say 'might' because this effort would have to be seen in light of: 1) the quality of students' performance on the generative activity, 2) the design of the generative activity, and 3) the possibility that engaging in the activity contributes to learning (i.e., future performance).

First, ratings of invested mental effort (or indicators like response times) cannot be meaningfully interpreted without looking at the quality (e.g., accuracy/errors made) of the associated performance (Van Gog & Paas, 2008). If a student has a high skill level and performed the generative activity correctly and with little effort, the amount of effort invested reflects the quality of the situation model and using effort as cue for the monitoring judgment ("this activity required little effort, so I expect to do well on the test") could contribute to higher monitoring accuracy. If, however, a student invested little effort but also performed poorly (e.g., made many commission errors), the low effort says nothing about the quality of the situation model and using it as a cue for a monitoring judgment ("this activity required little effort, so I expect to do well on the test") will lead to poor monitoring accuracy.

Second, the design of the generative activity may affect how informative invested mental effort is regarding the quality of the mental model (and hence, how useful it is as a cue). That is, we need to consider the *origin* of the effort (i.e., the processes effort is invested in). If the level of effort required to complete the generative activity is mainly dependent on students' situation model, then effort would be a diagnostic cue and could be expected to lead to higher monitoring accuracy. If, however, the level of effort required is affected by the way in which the generative activity is designed, then it is not necessarily predictive of students' understanding (cf. the concept of 'extraneous load'; Sweller, Ayres, & Kalyuga, 2011). For instance, if a concept mapping task is done on the computer and the program that is used is not very user-friendly, students would have to spend much effort on selecting nodes and arrows et cetera to get a visual representation of the concept map that they have in mind. While this would increase the effort they have to invest in the generative activity considerably, this is not very informative of the quality of their situation model, and thus, not diagnostic of their test performance.

Last, the usefulness of effort as a cue may be influenced by students experiencing to learn from generative activities (for which there is evidence, cf. Fiorella & Mayer, 2016). If generative activities requires high effort, but simultaneously contribute to improving the situation model (cf. the concept of germane cognitive load; Sweller et al., 2011), then students might underestimate their future test performance when using effort invested in the generative activity as a cue (unless they recognize that the high effort they invest is contributing to their learning; cf., goal-driven interpretation of effort; Koriat et al., 2014).

Up until now, we know of only two studies on mapping and drawing that incorporated mental effort measurements. In the study by Kostons and De Koning (2017), mental effort was measured to examine how drawing affects cognitive load and potentially cause overload. That is, mental effort was rated at the end of the entire trial (after the post-test) using an adaptation of the Paas' (1992) rating scale ("How much effort was required to solve the task?"). Students in the long-drawing condition reported higher levels of mental effort than in the other conditions, indicating that drawing contributes to experienced cognitive load; at least when performed during the entire learning task. Mental effort did not seem to be used as a cue by students for their monitoring judgments as there was no relation between students' monitoring judgments and their mental effort ratings. However, mental effort did seem to be used to some extent as a cue for their regulation judgments, indicated by a moderate correlation between mental effort ratings and restudy decisions ($r = 0.25$).

Schleinschok et al. (2017) also investigated the effect of drawing on mental effort. They adapted Kalyuga, Chandler, and Sweller's (1999) effort rating (i.e., "How easy or difficult was it to learn something about this phenomenon?"). Judgments of learning were made after each paragraph, while mental effort was measured at the end of the learning task. Contrary to Kostons and De Koning (2017) no difference was found in mental effort between those who did the drawing task and those who did not. The higher experienced mental effort in the long-drawing condition in Kostons and De Koning (2017) may be due to the fact that in that condition, students drew and used their drawings (and were allowed to adapt them) all along the trial, including the test, whereas in Schleinschok et al. (2017), students only drew after reading a paragraph. Future research should thus be cautious with allowing students to use and adapt drawings throughout the trial, because it can result in higher experienced cognitive load without improving monitoring accuracy or learning. In Schleinschok et al. (2017) effort ratings and monitoring judgments were negatively correlated ($r_{overall} = -0.38$), indicating that when students perceived the learning task as more effortful, they expected their test performance to be lower. This study indicates that, during a drawing intervention, mental effort is to a certain extent related to judgments of learning and thus that mental effort is potentially used as a cue when making judgments. Although it was not measured directly how and whether students used effort as a cue for monitoring in these studies, these findings call for further research on this topic where students, for example, are inquired about their effort interpretations specifically for the generative activities and it is measured how these affect monitoring, regulation, and actual learning.

Directions for Future Research

Research using the generation paradigm with mapping and drawing has yielded valuable information about how students' and teachers' cue utilization affected their monitoring and regulation accuracy. Yet, using the generation paradigm in mapping and drawing monitoring research offers unique opportunities for future research on further developing the measurement of cue diagnosticity and cue utilization and investigations of effort as a cue. These avenues for future research are described here.

Cue Diagnosticity Many studies seem to assume that performance cues (e.g., correct relations in a diagram) are more diagnostic than student cues (e.g., students' interest in the text topic) or task cues (e.g., text length) (e.g., Van de Pol et al., 2019). The studies that *have* measured cue diagnosticity mainly focused on the diagnosticity of performance cues. One exception is the study by Authors (SubmittedB) (n.d.), which measured the diagnosticity of a wide variety of performance, student, and task cues. However, for each of the 28 cues measured in the study, one diagnosticity index was computed for the whole sample, so that a cue (e.g., number of correct relations) always had the same diagnosticity for all students in the sample.

Yet, results of Authors (2020) suggest that the diagnosticity of a cue may not be the same for all students (also see supplementary material). For instance, the diagnosticity of diagram cues appeared to differ for students with high versus low absolute monitoring accuracy. That is, the number of completed boxes and omissions were more diagnostic for students with high monitoring accuracy than for students with low monitoring accuracy. This may be associated with differences in the cue manifestations, as the analyses indicated that the high monitoring accuracy group completed more boxes correctly, while also making more omission errors and

less commission errors than the low monitoring accuracy group. The number of completed boxes presumably is a more diagnostic cue for students high in monitoring accuracy because they completed more boxes *correctly*. That is, the information in the diagram (the number of completed boxes) relates to students' actual understanding (as measured with a performance test), which makes this cue diagnostic. In contrast, the low monitoring accuracy group left fewer boxes empty and made more mistakes in completing boxes (i.e., commission errors). These students' diagrams thus contain much information, but that information is often incorrect. When students, however, think that this information is correct and use this information to base their judgment upon, their judgment is likely not accurate.

Although individual differences in cue diagnosticity complicate things further for teachers and for students, it is important to take individual differences in cue diagnosticity into account in future research. It is advisable to employ designs that allow the examination of potentially differing effects between groups of students with high and low monitoring accuracy. Also, interventions to improve monitoring accuracy by focusing students' attention on diagnostic cues should probably take into account interindividual differences between students (e.g., instructing students to focus on empty boxes may not help those who leave few boxes empty yet make many commission errors). To account for the unproductive effect of commission errors, which are presumably one of the symptoms of the observed inter-individual differences, further research could consider the use of feedback to make students aware of their errors and therewith improve their monitoring accuracy.

To be able to measure the diagnosticity of a cue, one needs to measure both the actual manifestations of the cue and students' test scores. If one, for example, wishes to determine the diagnosticity of the cue "the number of correct relations in the diagram," one needs to objectively score the number of correct relations that are present in students' diagrams and relate this to students' test scores. If one wants to be able to use diagnosticity scores that differ between students, one should use several measurements of the same cue within each student (e.g., score the number of correct relations in all six diagrams for one student and relate this to the test scores on the six test questions).

Cue Utilization Mapping and drawing lend itself well for coding several cues such as correct relations or commission errors in students' diagrams (Van de Pol et al., 2019; Van Loon et al., 2014), paragraph use or number of nodes in concept maps (Wiley, 2019), or correct elements or detail level in drawings (Schleinschok et al., 2017; Thiede et al., 2019). Previous research on mapping and drawing has used the coded cues to infer cue utilization. Yet, this approach relies on the objective or actual cue manifestations (e.g., the actual number of commission errors), whereas students or teachers may not be aware of those objective cue manifestations (cf. Authors (SubmittedB), n.d.). The relation between a cue manifestation and the test score may therefore not represent actual cue utilization; rather, the relation between a person's *perception* of a cue manifestation and the person's test score represents the person's cue utilization.

Authors (SubmittedB) showed that, when aiming to acquire insight into whether cue utilization is actually based on diagnostic cues, it may be promising to ask teachers to also judge the actual manifestations of the cues they have used (e.g., "How many correct relations do you think the diagram contains?"). Those judgments of cue manifestations can then be related to text comprehension judgments. Previous research did find some relation between cue utilization based on actual cue manifestations and monitoring accuracy (e.g., Van Loon et al., 2014), indicating that it is likely that there is some relation between actual cue

manifestations and perceived cue values. Future research could further explore to what extent students' or teachers' cue perceptions differ from objective or actual cue manifestations. However, a downside of this method may be that participants need to first indicate what cues they have used and then also judge the cues, which may affect the judgment process. Although the study by Authors (SubmittedB) did not show differences between teachers' judgment accuracy when they were asked to only judge students' text comprehension versus when they were asked to judge students' text comprehension, indicate their cue utilization, and judge cue values, future research should further investigate whether asking participants what cues they used and to judge the cues would affect the judgment process.

Another measure of cue utilization that is sometimes used and that is useful for research on mapping and drawing is self-report (e.g., Bol, Riggs, Hacker, & Nunnery, 2010; Dinsmore & Parkinson, 2013; Hacker, Bol, & Bahbahani, 2008; Händel & Dresel, 2018; List & Alexander, 2015; Thiede et al., 2010; also see Hacker & Bol, 2019). These studies asked (mainly college) students where they based their judgments upon. Students' answers were coded into several cue categories such as social factors or task/test characteristics (Hacker et al., 2008). A more efficient way to measure self-reported cue utilization was used in Authors (SubmittedB), where teachers were asked to indicate on a list of 28 cues what cues they used for each judgment. Having the cue list available did not affect their judgment accuracy compared to having no cue list available. Future research should further investigate the promises and potential pitfalls of self-reported cue use. This could be done by, for example, comparing self-reported cue utilization with measures of cue utilization based on the correlation of people's judgments of cue manifestations and test score.

A possible downside of the self-report approach is that participants are asked to report cues in hindsight. A promising way to measure cue utilization *during* the monitoring process is using process measures. One such process measure is using concurrent think aloud methods, asking participants to self-report their cue use while monitoring (cf. Authors, SubmittedA; SubmittedB). These think-aloud protocols can be coded with regard to several cues. Moreover, the visual nature of the mapping and drawing tasks could make it particularly apt for process measures such as eye tracking, that can be used in combination with concurrent thinking aloud or cued retrospective reporting (verbal reports immediately after a task, based on a replay of the participants' eye movements, visualized as a circle or dot; Van Gog, Paas, Van Merriënboer, & Witte, 2005). Such a concurrent or retrospective think aloud procedure could be implemented during an extra inspection task, during which students return to their completed diagram/map/drawing when making monitoring judgments. Recordings of eye movements could reveal what diagram/map/drawing cues students (e.g., the time spent looking at completed boxes in a diagram, transitions between boxes in concept maps), and verbal reports would provide information on why and how they use this information. Moreover, differences in eye movements and verbal reports between students with high versus low monitoring accuracy could be used to unravel how inspection patterns are related to monitoring accuracy, which might aid the development of novel interventions.

Further, studies have focused on what cues students and teachers use for *monitoring*. However, it remains unknown what cues are used when making *regulation* decisions and allocating (re)study time. Van de Pol et al. (2019) have shown that diagramming had an effect on students' monitoring but not on their regulation and on teachers' regulation and not on their monitoring. People may therefore use different considerations and cues to arrive at their decisions (cf. Van de Pol et al., 2019). For instance, even when someone accurately monitors that a text is not yet well understood and that comprehension may benefit from restudy, the text

may not be selected again due to doubts about the likelihood of success. Further, when time constraints play a role, sometimes better-learned texts may be selected for restudy rather than less-learned texts (Kornell & Metcalfe, 2006).

Effort as a Cue for Monitoring and Regulation of Text Comprehension Using Mapping and Drawing Research using mapping and drawing as generative activities has the potential to enhance the use of diagnostic cues when monitoring and regulating learning of texts. Schleinschok et al. (2017) showed that students use effort as a cue, in such a way that a high level of effort is related to lower judgments of learning. Additionally, Kostons and De Koning (2017) showed that a higher level of effort is related to a lower number of texts selected for restudy. As explained earlier, whether the amount of invested effort is a diagnostic depends on (1) how the student performs on the generative activity, (2) the design of the generative activity, and (3) whether or not students experience to learn from the generative activity (i.e., further build their situation model). This leads to several issues regarding the use(fulness) of effort as a cue that future research would need to address.

First, students probably do not know whether they performed well on the generative activity, given that their monitoring accuracy is generally low. If students experienced little effort but their generative activity is of low quality, effort is probably not diagnostic and should thus not be used. Future research should focus on ways that help students to assess their work *and* to teach students when using effort as a cue is a good idea. A promising avenue to give students insight into the quality of their product may be to combine the mapping or drawing task with the idea unit standard procedure (Dunlosky, Hartwig, Rawson, and Lipko, 2010). This would entail having students compare their diagrams, drawings, or concept maps bit-by-bit to a standard (e.g., a correct diagram). Students are instructed to compare each bit of the standard (e.g., each diagram box) to their own diagram and determine whether they had this bit correct in their own work.

Second, students are probably not able to distinguish between effort that has its origin in the (poor) design of generative activities and effort that has its origin in the quality of the situation model. Only effort that has its origin in the quality of the situation model is probably diagnostic of students' understanding, not effort originating from a poor design. Therefore, future research on mental effort and generative activities should thoroughly check whether there is extraneous processing involved in completing the activity and minimize this extraneous load (cf. research question 3 of the EMR model). When extraneous load is minimized, students' mental effort will mainly originate in students' situation model and it can then be used as a (diagnostic) cue.

Third, engaging in generative activities can require effort (e.g., Kostons & De Koning, 2017) and students tend to interpret high effort as an indication of *poor* performance. However, the generative activities might also contribute to learning outcomes. When students experience to learn from the generative activity (i.e., experiencing that it helps further build their situation model), they may conclude that the amount of effort put in is a good predictor of the quality of their mental representation of the text. That is, they interpret the higher effort invested in the generative activity as contributing to a higher quality situation model, and, as such, high effort is diagnostic cue for monitoring. This also relates to cognitive psychological research on learning paired associates that has shown that experienced effort drives students' judgments of learning (Koriat, Nussinson, & Ackerman, 2014), suggesting that students can adopt either a data-driven interpretation of effort (i.e., experiential: higher experienced effort is interpreted as low learning) or a goal-driven interpretation (i.e., investment-driven: higher invested effort is

interpreted as high learning). This research shows that students are more inclined to interpret effort in a data-driven manner, but how this applies when generative activities are inserted before monitoring and to what extent this can aid in recognizing that high effort can also indicate high learning is unknown. It would be relevant to examine whether and when students experience the generative activity as a learning task to improve their situation model and whether this affects the relation between mental effort and monitoring judgments.

Future research on the relation between mental effort, students' monitoring, regulation, and learning, should carefully consider how the mental effort rating is included in the study design. We propose to add effort ratings in a similar way as judgments of learning or comprehension ratings: per learning task (e.g., text), and after the generative activity per learning task. Phrasing of the effort ratings should be designed with caution, as it may affect absolute level of ratings. Students are inclined to interpret effort as experienced effort (as a result of the learning task) but may be shifted to rate it as effort they decided to invest top-down depending on whether a learning goal was set (Koriat et al., 2014; Paas, 1992). Unknown is to what extent sequencing of the effort ratings and judgments matters. Typically, effort ratings are asked before judgments, but experimental research is needed to determine whether there are sequencing effects requiring attention in study design.

When it comes to regulation of learning, effort is a potentially highly important cue used by learners. That is, when deciding what to restudy it is likely that learners not only consider what texts were least understood, but also how much effort it takes to restudy these texts (also see Van Gog, Hoogerheide, and Van Harsel, n.d., this special issue). Learners possibly make a judgment of how much effort it will cost to restudy the text and to what extent this effort is likely to produce a learning result. Alternatively, they set a boundary on the number of texts they choose to restudy to limit effort, even when restudying all texts would be advisable based on mapping or drawing cues. This is different from the typical mental effort ratings where effort *previously invested in a task* is measured. During regulation, the amount of effort willing to invest in restudy is used in a goal-driven or top-down manner, and possibly interacts with the use of data-driven or bottom-up cues like the number of correct relations in the diagram that indicate which texts require effort to restudy.

To ensure that restudy is mostly based on text comprehension cues and less on which texts require least effort, it may be promising to combine the mapping or drawing task with the idea unit standard procedure (Dunlosky et al., 2010), with the aim to make the comprehension cues even more explicit and thereby ensure that effort is invested into the texts that need it most. All in all, it is time to start examining how mental effort, both experienced effort *and* the willingness to invest effort in restudy or future learning, can be validly measured and how this affects students monitoring and regulation of text learning.

Conclusion

Research on mapping and drawing has shown that using diagnostic cues, which give insight into the quality of students' understanding of relations in a text, seems to help students to arrive at more accurate monitoring judgments. Also, for teachers, diagram cues which became visible through inspection of students' maps or drawings, proved valuable to improve teachers' monitoring accuracy of students' understanding, but only when teachers had accurate insight into the actual quality of the generated info. Thus, mapping and drawing seems promising as a basis for future research aiming to acquire understanding about how cues are used when

making monitoring judgments, and to what extent these cues are valid indicators of performance.

Mapping and drawing offers unique opportunities for future research on advancing the measurement of cue diagnosticity and cue utilization and studying how one especially important cue, namely effort, is used as a cue for monitoring and regulation. Research should extend their focus on monitoring to regulation, and also include actual measures of dynamic, online monitoring and regulation of learning. Improving measures of cue diagnosticity and cue utilization can provide us with better understanding of how students and teachers monitor and regulate learning, which can help design effective interventions to foster these important skills.

Funding Information This work was funded by a Veni grant from the Netherlands Organization for Scientific Research awarded to the first author (grant number: 451-16-012).

References

Authors (2020). Diagnosticity of diagram cues for text comprehension: Differences for high/low monitoring accuracy. Paper accepted for presentation at the EARLI SIG 6&8 conference, Dresden, Germany.

Anderson, M. C., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica, 128*(1), 110–118. https://doi.org/10.1016/j.actpsy.2007.10.006.

Baars, M., Van Gog, T., de Bruin, A., & Paas, F. (2014a). Effects of problem solving after worked example study on primary school children's monitoring accuracy. *Applied Cognitive Psychology, 28*(3), 382–391. https://doi.org/10.1002/acp.3008.

Baars, M., Van Gog, T., de Bruin, A., & Paas, F. (2018). Accuracy of primary school children's immediate and delayed judgments of learning about problem-solving tasks. *Studies in Educational Evaluation, 58*, 51–59. https://doi.org/10.1016/j.stueduc.2018.05.010.

Baars, M., Vink, S., van Gog, T., de Bruin, A., & Paas, F. (2014b). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learning and Instruction, 33*, 92–107. https://doi.org/10.1016/j.learninstruc.2014.04.004.

Baars, M., Visser, S., Van Gog, T., de Bruin, A., & Paas, F. (2013). Completion of partially worked-out examples as a generation strategy for improving monitoring accuracy. *Contemporary Educational Psychology, 38*(4), 395–406. https://doi.org/10.1016/j.cedpsych.2013.09.001.

Baars, M., Wijnia, L., De Bruin, A., & Paas, F. (This special issue). The relation between student's effort and monitoring judgments during learning: a meta-analysis.

Biancarosa, G., & Snow, C. E. (2004). Reading next: A vision for action and research in middle and high school literacy: A report from Carnegie Corporation of New York. Alliance for Excellent Education.

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

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.

De Bruin, A. B. H., Roelle, J., & Baars, M. (This special issue). Synthesizing cognitive load and self-regulation theory: A theoretical framework and research agenda.

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

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.

Dunlosky, J., & Ariel, R. (2011). The influence of agenda-based and habitual processes on item selection during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(4), 899–912. https://doi.org/10.1037/a0023064.

Dunlosky, J., Hartwig, M. K., Rawson, K. A., & Lipko, A. R. (2010). Improving college students' evaluation of text learning using idea-unit standards. *The Quarterly Journal of Experimental Psychology, 64*(3), 467–484. https://doi.org/10.1080/17470218.2010.502239.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003.

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.

Foster, N. L., Was, C. A., Dunlosky, J., & Isaacson, R. M. (2017). Even after thirteen class exams, students are still overconfident: The role of memory for past exam performance in student predictions. *Metacognition and Learning, 12*(1), 1–19. https://doi.org/10.1007/s11409-016-9158-6.

Ghetti, S., Hembacher, E., & Coughlin, C. A. (2013). Feeling uncertain and acting on it during the preschool years: A metacognitive approach. *Child Development Perspectives, 7*(3), 160–165. https://doi.org/10.1111/cdep.12035.

Griffin, T., Mielicki, M., & Wiley, J. (2019). Improving students' metacomprehension accuracy. In J. Dunlosky & K. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 619–646). Cambridge: Cambridge University Press.

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

Hacker, D. J., & Bol, L. (2019). Calibration and self-regulated learning: Making the connections. In J. Dunlosky & K. A. Rawson (Eds.), *Cambridge handbook of cognition and education* (pp. 647–677). NY, Cambridge: New York.

Händel, M., & Dresel, M. (2018). Confidence in performance judgment accuracy: The unskilled and unaware effect revisited. *Metacognition and Learning, 13*(3), 265–285. https://doi.org/10.1007/s11409-018-9185-6.

Kalyuga, S. (2006). Assessment of learners' organised knowledge structures in adaptive learning environments. *Applied Cognitive Psychology, 20*(3), 333–342. https://doi.org/10.1002/acp.1249.

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology, 13*(4), 351–371. https://doi.org/10.1002/(SICI)1099-0720(199908)13:4<351::AID-ACP589>3.0.CO;2-6.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49*(4), 294–303. https://doi.org/10.1037/0003-066X.49.4.294.

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.

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.

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52*(4), 478–492. https://doi.org/10.1016/j.jml.2005.01.001.

Koriat, A., Nussinson, R., & Ackerman, R. (2014). Judgments of learning depend on how learners interpret study effort. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(6), 1624–1637. https://doi.org/10.1037/xlm0000009.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(3), 609–622. https://doi.org/10.1037/0278-7393.32.3.609.

List, A., & Alexander, P. A. (2015). Examining response confidence in multiple text tasks. *Metacognition and Learning, 10*(3), 407–436. https://doi.org/10.1007/s11409-015-9138-2.

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.

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.

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*(4), 429–434. https://doi.org/10.1037/0022-0663.84.4.429.

Redford, J. S., Thiede, K. W., Wiley, J., & Griffin, T. D. (2012). Concept mapping improves metacomprehension accuracy among 7th graders. *Learning and Instruction, 22*(4), 262–270. https://doi.org/10.1016/j.learninstruc.2011.10.007.

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.

Smit, J., van Eerde, H., & Bakker, A. (2013). A conceptualisation of whole-class scaffolding. *British Educational Research Journal, 39*(5), 817–834. https://doi.org/10.1002/berj.3007.

Steiner, M., van Loon, M. H., Bayard, N. S., & Roebers, C. M. (2020). Development of children's monitoring and control when learning from texts: Effects of age and test format. *Metacognition and Learning, 15*(1), 3–27. https://doi.org/10.1007/s11409-019-09218-3.

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.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Intrinsic and extraneous cognitive load. In J. Sweller, P. Ayres, & S. Kalyuga (Eds.), *Cognitive load theory* (pp. 57–69). Basel: Springer.

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

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

Thiede, K., Oswalt, S., Brendefur, J., Carney, M., & Osguthorpe, R. (2019a). Teachers' judgments of student learning of mathematics. In J. Dunlosky & K. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 678–695). Cambridge: Cambridge University Press.

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes, 47*(4), 331–362. https://doi.org/10.1080/01638530902959927.

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. (2009). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 85–106). London: Routledge.

Thiede, K., Wright, K. L., Hagenah, S., & Wenner, J. (2019b). Drawings as diagnostic cues for metacomprehension judgment. In N. Feza (Ed.), Metacognition in learning. **InTechOpen**. https://doi.org/10.5772/intechopen.86959.

Van de Pol, J., de Bruin, A. B., van Loon, M. H., & van Gog, T. (2019). Students' and teachers' monitoring and regulation of students' text comprehension: Effects of comprehension cue availability. *Contemporary Educational Psychology, 56*, 236–249. https://doi.org/10.1016/j.cedpsych.2019.02.001.

Van de Pol, J., van den Boom-Muilenburg, S. N., & van Gog, T. (SubmittedA). Exploring the relations between teachers' cue-utilization, monitoring and regulation of students' text learning.

Van de Pol, J., van Gog, T., & Thiede, K. Ignoring non-diagnostic cues and correctly interpreting diagnostic cues improves teachers' monitoring accuracy of students' text comprehension

Van Gog, T., Hoogerheide, V., & Van Harsel, M. (This special issue). The role of mental effort in fostering self-regulated learning with problem-solving tasks.

Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist, 43*(1), 16–26. https://doi.org/10.1080/00461520701756248.

Van Gog, T., Paas, F., Van Merriënboer, J. J., & Witte, P. (2005). Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology: Applied, 11*(4), 237–244. https://doi.org/10.1037/1076-898X.11.4.237.

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.

Wiley, J. (2019). Picture this! Effects of photographs, diagrams, animations, and sketching on learning and beliefs about learning from a geoscience text. *Applied Cognitive Psychology, 33*(1), 9–19. https://doi.org/10.1002/acp.3495.

Wiley, J., Thiede, K. W., & Griffin, T. D. (2016). Improving metacomprehension with the situation-model approach. In K. Mokhtari (Ed.), *Improving reading comprehension through metacognitive reading instruction for first and second language readers* (pp. 93–110). Lanham, MD: Rowman & Littlefield.

Wittwer, J., & Renkl, A. (2008). Why instructional explanations often do not work: A framework for understanding the effectiveness of instructional explanations. *Educational Psychologist, 43*(1), 49–64. https://doi.org/10.1080/00461520701756420.

Affiliations

Janneke van de Pol[1] · Mariëtte van Loon[2] · Tamara van Gog[1] · Sophia Braumann[1] · Anique de Bruin[3]

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

[1] Department of Education, Utrecht University, PO Box 80.140, 3508 TC Utrecht, The Netherlands

[2] School of Health Professions Education, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands

[3] Department of Psychology, University of Bern, Fabrikstrasse 8, 3012 Bern, Switzerland