Contemporary Educational Psychology 56 (2019) 236–249

Contents lists available at ScienceDirect

Contemporary Educational Psychology

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

Students' and teachers' monitoring and regulation of students' text comprehension: Effects of comprehension cue availability

Janneke van de Pol[a,*], Anique B.H. de Bruin[b], Mariëtte H. van Loon[c], Tamara van Gog[a]

[a] Department of Education, Utrecht University, PO Box 80.140, 3508 TC Utrecht, the Netherlands
[b] School of Health Professions Education, Maastricht University, P.O. Box 616, 6200 MD Maastricht, the Netherlands
[c] Department of Psychology, University of Bern, Fabrikstrasse 8, 3012 Bern, Switzerland

ARTICLE INFO

Keywords:
Monitoring accuracy
Regulation accuracy
Metacognition
Self-regulated learning
Teacher regulation

ABSTRACT

For regulation of text learning to be effective, students need to accurately monitor their text comprehension. Similarly, to provide adaptive instruction, teachers need to accurately monitor and regulate students' text comprehension. Performing generative activities prior to monitoring has been suggested to provide students with diagnostic cues, improving monitoring accuracy; an open question is whether this would also help teachers. We investigated whether two generative activities, diagram completion and diagram drawing, improved secondary education students' (n = 248) monitoring and regulation accuracy of text comprehension (Experiment 1) and whether viewing students' diagrams improved *teachers'* (N = 18) monitoring and regulation of students' text comprehension (Experiment 2). Students' monitoring and teachers' regulation accuracy was higher in the diagramming conditions than in the no-diagramming condition. Students *and* teachers used diagnostic cues when judging students' text comprehension: Improving students' monitoring and teachers' regulation of students' text comprehension relies on improving accessibility of diagnostic cues.

1. Introduction

Students' monitoring of how well they comprehend study materials is crucial for their study behavior and their academic success (Dunlosky & Rawson, 2012; Thiede, Anderson, & Therriault, 2003; Winne & Hadwin, 1998). In addition, *teachers'* monitoring of their students' comprehension is pivotal in the teaching and learning process. The accuracy of teachers' monitoring affects the instructional quality, which in turn affects student achievement (Südkamp, Kaiser, & Möller, 2012). A review has shown that instructional quality is high whena teacher's instruction is adapted to a student's comprehension (Van de Pol, Volman, & Beishuizen, 2010). Thus, both student and teacher monitoring should be as accurate as possible to optimize students' and teachers' regulation[1] (or guidance) of students' learning (Pino-Pasternak, Tolmie, & Whitebread, 2010; Wood, Bruner, & Ross, 1976).

Although monitoring accuracy of text comprehension is often low, research on students has shown that monitoring accuracy improves when performing 'generative activities' (see Fiorella & Mayer, 2016) prior to making monitoring judgments about texts (e.g., Van Loon, de Bruin, van Gog, van Merriënboer, & Dunlosky, 2014). Generative activities refer to activities that help students to make their text comprehension explicit by actively generating (i.e., retrieving) textual information from memory (Thiede & De Bruin, 2017), e.g., by generating summaries (Thiede & Anderson, 2003), or completing diagrams of causal relations (Van Loon et al., 2014). Generative activities are held to improve the accuracy of monitoring judgments because they provide students with cues that are diagnostic (i.e., predictive of high test scores) of their text comprehension (cf. the cue-utilization framework by Koriat, 1997). Recent research suggests that teachers' cue use also seems to affect their monitoring accuracy (Thiede et al., 2015), yet it is an open question whether cues arising from students' generative activities would also improve the accuracy of teachers' monitoring of their students' comprehension. Therefore, we aim to investigate in the present study whether two generative activities (diagram completion and diagram drawing) performed by students, would improve both students' (Experiment 1) and teachers' (Experiment 2) monitoring of students' text comprehension and subsequent study regulation with regard to text learning.

2. Students' monitoring and regulation accuracy

The accuracy of students' monitoring of their own text comprehension is often low. Correlations between judgments of comprehension and

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

[1] Following Vermunt and Verloop (1999) we define teacher regulation as the steering of students' learning process.

https://doi.org/10.1016/j.cedpsych.2019.02.001

Available online 07 February 2019
0361-476X/

objective indicators of comprehension (e.g., exam scores) typically do not exceed .27[2] (Dunlosky & Lipko, 2007). This is attributed to students' use of low quality or non-diagnostic cues when making monitoring judgments (Dunlosky, Mueller, & Thiede, 2014). That is, according to the cue-utilization framework (Koriat, 1997), monitoring is inferential in nature. People do not have direct access to the quality of their cognitive states, but have to infer their level of comprehension and knowledge based on other information or 'cues' that are available. The extent to which these cues are predictive, or diagnostic, of actual comprehension determines the quality of the cues. Specifically, the use of cues that are highly diagnostic of the quality of comprehension (e.g., how well one can summarize the gist of the text) results in more accurate monitoring judgments than the use of cues that have low diagnostic value for actual comprehension (e.g., how well one likes the topic of the text). Therefore, the key to improving monitoring accuracy seems to lie in helping students focus on diagnostic cues.

One way to do so is to have students engage in so-called 'generative activities' (see Fiorella & Mayer, 2016). Generative activities that have proven helpful for improving the accuracy of monitoring judgments of text comprehension are generating summaries (Thiede & Anderson, 2003), keywords (De Bruin, Thiede, Camp, & Redford, 2011; Thiede, Dunlosky, Griffin, & Wiley, 2005), drawings (Kostons & de Koning, 2017; Schleinschok, Eitel, & Scheiter, 2017), concept maps (Redford, Thiede, Wiley, & Griffin, 2012), and completing diagrams (Van Loon et al., 2014) of the studied texts prior to judging comprehension of the texts. Similarly, when acquiring problem-solving skills by means of studying worked examples (i.e., examples that provide a step-by-step demonstration of how the problem is solved), having students generate some (i.e., example completion; Baars, Visser, van Gog, de Bruin, & Paas, 2013) or all (i.e., practice problem solving after example study; Baars, Van Gog, De Bruin, & Paas, 2014, 2017) problem-solving steps by themselves, has been shown to improve monitoring accuracy compared to studying the worked examples only without generating (some or all) steps.

The present study is concerned with monitoring accuracy during text comprehension. Presumably, engaging in the abovementioned generative activities provides students with cues regarding the quality of their comprehension of the gist of a text. For instance, completing diagrams about texts that contain temporal causal relations (Van Loon et al., 2014) provides students, amongst others, with cues about which relations represented in the diagram they could or could not complete, which is indicative of their text comprehension. It is assumed that engaging in generative activities gives students insight into the quality of their situation model. A situation model can be characterized as a deep level of text understanding (i.e., beyond verbatim/semantic features) at which the situation described in the text is organized into a coherent mental representation and integrated with prior knowledge (Kintsch, 1988; Thiede & Anderson, 2003). Importantly, more accurate monitoring has been shown to result in more effective or accurate study regulation (e.g., selecting those texts for restudy that are least well understood) in some studies (De Bruin et al., 2011; Kostons & de Koning, 2017; Mihalca, Mengelkamp, & Schnotz, 2017; Van Loon et al., 2014). Note that some of those generative activities (e.g., Van Loon et al., 2014) are only effective for improving monitoring accuracy when they are performed at a delay after text study. When they are performed immediately after studying a text, information from the text is still available in working memory, which may ease the process of generation, but the cues that are gained from this experience are not necessarily predictive of long-term memory of the gist of the text. When the generative activities are performed at a delay, information from working memory has decayed and students have to rely on long-term memory. The cues gained from this experience, which is more similar to the later test situation, will be more predictive of their later comprehension test performance, and will therefore improve monitoring accuracy (Thiede et al., 2005). Other activities are, however, also effective at providing students with diagnostic cues regarding their comprehension when performed during or immediately after text study (e.g., drawing: Kostons & de Koning, 2017; Schleinschok et al., 2017; or concept maps: Redford et al., 2012).

Van Loon et al. (2014) showed that the delayed completion of diagrams about causal relations in texts provided students with diagnostic cues of their comprehension of causal relations, but not of comprehension of facts. In this study of Van Loon et al. (2014), 15-year-old students were presented with six texts containing several cause-and-effect relations that they had to learn. Students who completed blank, pre-printed diagrams of the causal relations (see Fig. 1) before judging their text comprehension had higher monitoring accuracy of their text comprehension (hereafter referred to as monitoring accuracy for simplicity) than students who did not complete diagrams (mean gamma correlation between judgments and test performance in diagrams condition 0.56; in no diagrams condition 0.07). Monitoring accuracy of students who completed diagrams immediately after text study fell in between (mean gamma correlation was 0.28; this did not differ significantly from either the delayed or the no completion condition). Analysis of the content of the diagrams suggested that diagram completion indeed provided students with access to, and stimulated their use of diagnostic cues. For instance, the number of correct relations they generated and the number of boxes they left blank correlated with their monitoring judgments, which suggests that students used this information as a basis for their judgments. As these cues were also diagnostic of test performance (and more so in the delayed completion condition), students' monitoring accuracy improved.

One interesting question raised by this study, is whether students' monitoring accuracy improved from engaging in relation generation as such, or was aided by the support that the pre-printed blank diagram boxes provided. Asking students to draw the diagrams themselves would on the one hand require more generative processing than diagram completion (e.g., deciding how different elements relate to each other) and might therefore provide more comprehension cues (cf. studies on free drawing; e.g., Redford et al., 2012). On the other hand, students may be more likely to generate information that is irrelevant for the test or even completely incorrect (Finn & Tauber, 2015) when they have to generate the entire diagram themselves. If they would then use the mere act of retrieving from memory (i.e., accessibility) or the fluency with which they retrieved as a cue for comprehension, regardless of the actual quality of the retrieved information, their monitoring accuracy will be hampered. This question of whether or not diagram *drawing* would be less, more, or equally effective than diagram completion is also of interest to educational practice, as the former would be easier to implement than the latter, which requires teachers or instructional designers to prepare pre-printed diagrams for each text.

We addressed this question in Experiment 1, by comparing the effect of delayed diagram completion, delayed diagram drawing, or no generative activity on *students'* monitoring and regulation accuracy when studying causal relations texts. In Experiment 2, we investigated whether viewing students' products of these generative activities (i.e., completed diagrams or drawn diagrams), would also improve the accuracy of *teachers'* monitoring of their students' text comprehension and teachers' regulation accuracy.

3. Teachers' monitoring and regulation accuracy

Research on teachers' monitoring of their students' performance has mostly focused on observing or describing the relation between teachers' judgments and students' test performance, using a rank correlation, indicating the degree to which teachers are able to accurately rank their students according to performance (Cronbach, 1955, also see

[2] Expressed as a gamma correlation (Nelson, 1984), which indicates to what extent students can discriminate between well studied materials and less studied materials. Gamma correlations range from −1 to +1. A gamma of +1 indicates perfect monitoring accuracy (perfect discrimination).

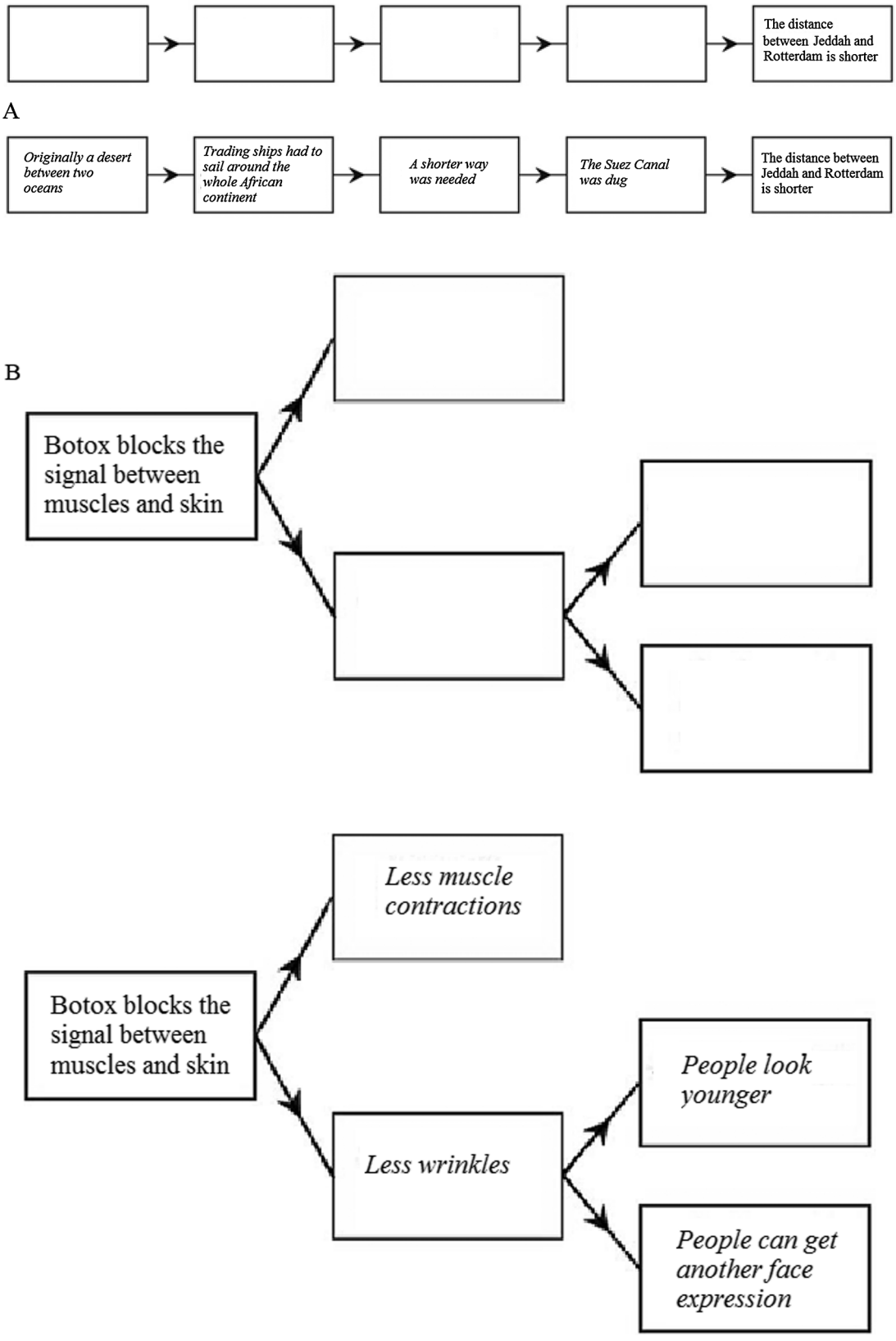

Fig. 1. An empty and a correctly completed diagram for the text 'Suez Canal' (A). An empty and a correctly completed diagram for the text 'Botox' (B). Reprinted from "Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy," by Van Loon et al. (2014), Acta Psychologica, 151, p. 145. Copyright 2014 by M van Loon. Reprinted with permission.

Südkamp, Möller, & Pohlmann, 2008). In studies using this measure, teacher monitoring accuracy appears to be relatively high: A meta-analysis of 75 studies on teachers' monitoring accuracy of students' academic achievement showed a mean correlation of 0.63 between teachers' judgments and students' achievement (Südkamp et al., 2012). Yet, there is great variance between teachers and much room for

improvement in accuracy. Improving teachers' monitoring accuracy is important as it impacts the accuracy of their regulation, that is, the adaptivity of their instructional support to an individual student's needs (Behrmann & Souvignier, 2015; Karing, Pfost, & Artelt, 2011; Herppich et al., in press; Klug, Bruder, Kelava, Spiel, & Schmitz, 2013).

In contrast to the question of how to improve students' monitoring of their own comprehension by directing their attention to diagnostic cues, intervention research to improve teachers' monitoring accuracy of their students' comprehension is scarce. In a study by Van de Pol, Volman, Oort, and Beishuizen (2014), social studies teachers in secondary education participated in a professional development program aimed at improving the teachers' ability to provide adaptive instructional support. Teachers were encouraged to ask students to explicitly demonstrate their comprehension in face-to-face conversations with the teachers. This provided teachers with more diagnostic information about students' comprehension, which resulted in more accurate adaptation of instructional support (i.e., higher regulation accuracy) compared to teachers who did not participate in this program. Yet, teachers' actual monitoring accuracy, which is hypothesized to have caused the increased regulation accuracy, was not measured in this study.

Thiede et al. (2015) conducted one of the first studies on teachers' monitoring accuracy that used an intervention *and* measured the accuracy of teachers' monitoring. Secondary mathematics teachers did or did not participate in a professional development program aimed at stimulating teachers to examine students' mathematical thinking, which is assumed to yield diagnostic cues about students' comprehension. Teachers who participated in the program were better able to accurately rank their students with regard to their conceptual mathematical comprehension (gamma correlation of 0.20) than teachers who did not participate (gamma correlation of 0.02). However, there was still much room for improvement given the relatively low gamma correlation. As suggested by Thiede et al. (2015) helping teachers to recognize and use diagnostic cues may further increase their monitoring accuracy and regulation of students' learning process.

Focusing teachers' attention on diagnostic cues thus seems to be key in promoting monitoring and regulation accuracy. In Experiment 2 of the present study, we investigated whether providing teachers with the results of generative activities performed by students in Experiment 1, that is, completed or drawn diagrams, would improve teachers' monitoring and regulation accuracy, compared to not having such information available. Viewing completed and drawn diagrams would provide teachers with access to diagnostic cues of students' comprehension of a text, which can be expected to improve their monitoring accuracy, and improved monitoring accuracy can in turn be expected to improve regulation accuracy. As for students, it is interesting to investigate whether or not there are differences between drawn and completed diagrams for teachers' monitoring and regulation accuracy. On the one hand, the support that is inherent in the pre-printed diagram boxes in the completion condition, might also aid teachers' monitoring accuracy for the same reasons it would aid students' monitoring (e.g., by showing at a glance what a student could not complete). On the other hand, drawn diagrams may contain more diagnostic cues for teachers (e.g., did the student comprehend the structure of the relations) and unlike students themselves, the teachers would not be hindered by non-diagnostic experiential cues (e.g., fluency).

We focus on teachers' *intra-individual* monitoring judgments (i.e., relative accuracy) about their students' comprehension, that is, the extent to which teachers know which materials students understand compared to other materials. Most research in the teacher monitoring literature has focused on more general measures of accuracy, for example, estimating each student's general level (e.g., Kaiser, Retelsdorf, Südkamp, & Möller, 2013; Kaiser, Möller, Helm, & Kunter, 2015) or ranking students within a class regarding their level (e.g., Thiede et al., 2015). Yet intra-individual monitoring accuracy is of importance for high quality teaching. Knowing what texts (or tasks) a student does or does not comprehend (or master) is prerequisite for teachers' regulation of students' learning process. Without accurate monitoring teachers cannot adapt instructional support to the student's level of comprehension (and adaptive instruction promotes students' learning; Van de Pol, Volman, Oort, & Beishuizen, 2015).

4. The present study

In the present study we compared the effect of a diagram completion, a diagram drawing, and a no-diagram control group on students' monitoring and regulation accuracy of their text comprehension (Experiment 1) and teachers' (Experiment 2) monitoring and regulation accuracy of their students' text comprehension. In Experiment 1, students first read six texts, completed a diagram task or a filler task, monitored their comprehension of the six texts (judging both comprehension of cause-and-effect relations and factual details, cf. Van Loon et al., 2014), and then selected texts for restudy (i.e., regulation) before completing a test on their knowledge of relations and facts.

Because the diagram completion and drawing tasks explicitly focused students' (and teachers') attention on their comprehension of the causal events in the texts rather than on the details, we hypothesized (cf. findings by Van Loon et al., 2014) that students in the diagram completion and diagram drawing conditions would show more accurate monitoring (H1.1) and regulation (H1.2) regarding their comprehension of cause-and-effect relations than the control group, but not regarding their comprehension of factual information (replication of Van Loon et al., 2014). Regarding differences between diagram completion and diagram drawing, one could on the one hand expect that drawing might lead to higher accuracy than completion, as it requires more generative processing than diagram completion and might therefore provide more comprehension cues, in which case monitoring and regulation would be expected to be more accurate in the diagram drawing condition compared to the diagram completion condition. On the other hand, completion might lead to higher accuracy than drawing, because diagram drawing might also yield more non-diagnostic cues than completing pre-printed diagrams, in which case monitoring and regulation would be expected to be more accurate in the diagram completion condition compared to the diagram drawing condition. Therefore, this is explored as an open question. Furthermore, to explore whether students used their monitoring judgments while making restudy selections, we computed the correlation between students' Judgments of Learning (JOLs) and restudy selections in each condition.

To acquire more insight into effects of the diagram tasks on monitoring accuracy, we explored whether the cues provided by the diagrams (i.e., number of commission errors -incorrect information provided in the diagram; omissions –missing information in the diagram; completed boxes; and correct relations) were predictive of students' actual performance (i.e., cue-diagnosticity), and whether students showed indications of use of these diagram cues in their judgments (i.e., cue-utilization), by correlating the diagram cues with their monitoring judgments. Finally, we investigated the relation between students' judgments and restudy selections and expected significant correlations between the two (Thiede, Redford, Wiley, & Griffin, 2017; Van Loon et al., 2014).

In Experiment 2, teachers judged students' comprehension of the six studied texts (i.e., monitoring) and selected texts that students should restudy (i.e., regulation), based on students' completed diagrams, drawn diagrams, or based on student names only (control condition). This enabled us to assess the added value of cues that could be inferred from seeing a students' work above and beyond the information they have about their student, for making monitoring and regulation judgments. This resembles the normal classroom situation; as teachers know their students, they always have access to student cues, and sometimes they additionally can consult a product the student created, when making instructional decisions.

We hypothesized that teachers who viewed students' diagrams (either completed or drawn) would show more accurate monitoring (H2.1) and regulation (H2.2) judgments regarding students'

comprehension of cause-and-effect relations, though not necessarily of facts. We additionally explored whether the accuracy of teachers' monitoring and regulation of students' comprehension of cause-and-effect relations would differ depending on the type of representation they viewed (completed or drawn diagrams); one could expect that viewing drawn diagrams might lead to higher monitoring and regulation accuracy than completion because drawings provide more (diagnostic) comprehension cues or to lower monitoring and regulation accuracy than completion, because drawing also provides more non-diagnostic cues. We also explored whether teachers used their monitoring judgments while making restudy selections (cf. Experiment 1).

Moreover, we explored whether teachers' judgments were related to certain diagram cues, which may indicate their cue-utilization (see Experiment 1) in their judgments (i.e., cue-utilization). In addition, we investigated the relation between teachers' judgments and restudy selections and expected significant correlations between the two (Thiede et al., 2017; Van Loon et al., 2014). Finally, we exploratively and descriptively compared students' and teachers' monitoring and regulation accuracy, as the effects of the diagram tasks may differ for students and teachers.

5. Experiment 1: Students' monitoring and regulation accuracy

5.1. Method

5.1.1. Participants

Participants were Dutch secondary education students. Students and their parents/caregivers were informed about the study procedure; participation was voluntary and students could withdraw at any moment. All students in the 18 classrooms from 11 schools were tested resulting in a total number of 426 participants. Because it would be too time consuming for the teachers to make judgments on all of their students in Experiment 2, they made judgments about 15 students; 5 students were randomly selected from each of the three conditions from Experiment 1. This sub-selection of participants consisted of 248 students (59% girl; 96% Dutch; M_{age} = 14.60, SD = 0.63; 12.1% had dyslexia), of whom 82 were enrolled in senior general secondary education (the second highest level of secondary education in the Netherlands), and 166 in pre-university education (the highest level of secondary education in the Netherlands).[3] We complied with the APA ethical standards for treatment of human participants, informed consent, and data management.

5.1.2. Research design

Experiment 1 had a between-subjects design; participants were randomly assigned to the control condition (n = 84), diagram completion condition (n = 85), or diagram drawing condition (n = 79). The only difference between the conditions was the experimental task. There were no significant differences between conditions regarding students' gender, age, nationality, year, dyslexia and education level, all $ps > 0.08$.

5.1.3. Materials

We presented the materials in six paper booklets representing the six phases of Experiment: (1) Practice materials, (2) Texts, (3) Diagram (or filler) task, (4) JOLs, (5) Restudy selections, (6) Test. Six versions were used that differed with regard to the order in which the six texts were presented, using a Latin Square Design. Within each version (i.e., the materials for one student), the order of the texts was constant across the booklets (e.g., when making judgments, restudy selections, et cetera). All materials were presented in Dutch, and presentation was self-paced. When starting and finishing with each booklet, students were asked to fill out the start- and end-time of working with the booklet.

Booklet 1: Practice materials. This booklet contained a page on which participants could write their names and demographic information. Further, the participants were presented with two example texts (text 1 about the heart and text 2 about suburbs) and example questions about facts and causal relations; the texts and questions were similar to the examples used in the study by Van Loon et al. (2014). After example text 1, all students filled out a prestructured blank diagram about text 1 so they could practice the diagram completion task. After example text 2, students practiced drawing a diagram. That is, they were given the content of one text box and were instructed to draw and fill out the missing text boxes with regard to text 2. Further, the booklet contained information that, when finishing a booklet, the student could continue with the next booklet, but could not go back. When a student finished a booklet, it was collected by the experimenter.

Booklet 2: Texts. Study texts were similar to the texts used by Van Loon et al. (2014); the topics of the texts were "Botox", "Sinking of metro cars", "Concrete constructions", "Money does not bring happiness", "The Suez Canal", and "Music makes smart" (see Appendix A for example texts). All texts were presented in a single paragraph on a separate page. The average text length was 169.33 words (SD = 9.72). Each text contained exactly five clauses to convey causal relations. For three texts, causal relations were serial (e.g., the text "Suez"; each cause is followed by one effect only), the other three texts contained both serial and parallel causal relations (e.g., the text "Botox"; each cause can be followed by one or more effects).

Booklet 3: Diagram (or filler) task. Participants in the diagram completion condition received a booklet in which diagrams were printed, with each page containing one diagram. The title of the text was printed at the top of the page and, similar to the diagram completion condition by Van Loon et al. (2014), one of the diagram boxes was filled out. Participants had to complete the remaining diagram boxes.

In the diagram drawing condition, participants received a booklet in which each page showed the title of the text, followed by instructions that participants had to draw a diagram containing five boxes that could be either next to each other (serial) or below each other (parallel); a statement that needed to go into one of the boxes was given (this statement was similar to the one in the filled-out text box in the diagram completion condition).

The booklet for the control condition contained a filler task, which was a picture-matching task. These were also presented on six separate pages. The top of each page contained the title of the text, and the same statement was presented to the diagram completion and diagram drawing groups. The two pictures were related to the topics of the texts and participants were instructed to find four differences between the two pictures.

Booklet 4: JOLs. This booklet presented two JOL scales per text. On each page, participants saw the title of the text, and the question 'What percentage of questions do you expect to answer correctly for text [name of the text]?' which they answered separately for questions about facts ('Questions about facts:') and questions about cause-and-effect relations ('Questions about relations:'). These JOLs were made by indicating a value on a six-point scale ranging from 0% − 100%, in steps of 20%.

Booklet 5: Restudy selections. In this booklet, participants were instructed that they were about to take a test on which they had to show their understanding of the relations and memory of the facts in the texts. Participants were presented with the titles of the 6 texts listed below each other. They indicated with a check mark which text(s) they wanted to select for restudy ('Which text(s) do you want to read again before taking the test?'). They did, however, not get the opportunity to actually restudy these text(s), because this would affect their test performance (booklet 6) and would therefore interfere with analyses of JOL and restudy accuracy (Kimball & Metcalfe, 2003). After students had indicated which text(s) they wanted to restudy, they were debriefed that they were not going to restudy these texts, but that they were going straight to the test.

Booklet 6: Test. This booklet presented participants with test questions. Per text, participants answered one question about cause-and-effect relations, for which students could gain four points, one per relation (printed on the first page) and five questions about facts, for

[3] Test-scores for relations and facts were not significantly different between students at these two levels of education.

which students could gain five points (printed on a subsequent page) (see for example questions Appendix B). Participants were required to write down the answer to the questions in the text box that was printed below each question. On the bottom of each page, there was an instruction that upon turning the page, participants were not allowed to go back and change their answer to previous questions.

Puzzle booklet. A puzzle-booklet with Sudoku-puzzles and word-puzzles was provided after the test to keep participants occupied until all students in their classroom were ready with the experimental tasks.

5.2. Procedure

The students were tested in their own classroom. First, the students filled out some general information (gender, birth date, birth country of the student and his/her mother and father, and whether or not the student had dyslexia). Then, the experimenter instructed the students about the experiment and practiced with the students, using the practice booklet (booklet 1). They were then instructed that they would not receive any feedback during the experiment and that they were not allowed to go back to prior booklets or within booklets to prior texts. After that, students could start with the actual experiment, working through booklets 2–6 at their own pace. First, students read the six texts one by one (booklet 2). Then, depending on assigned condition, the students completed or drew pre-structured diagrams about each text or engaged in the picture-matching filler task (booklet 3). Subsequently, students made JOLs for each of the texts (booklet 4) and then indicated which text (s) they would like to read again before taking the test (booklet 5). Finally, the students completed the test (booklet 6). Students who finished early worked on the puzzle booklet to not disturb the others. Each time a student finished a booklet, he/she put the booklet upside down at the edge of the table and the experimenter would immediately collect the booklet. For each booklet, the students recorded the start and end time. The total duration of the experiment was approximately 60 min.

5.3. Scoring of responses

5.3.1. Scoring of test responses

Responses to the questions about causal relations and facts were scored in line with the scoring criteria by Van Loon et al. (2014). Causal relations scores indicate the number of correct causal relations in the response, ranging from 0 (no relations correct) to 4 (all relations correct) per text. Both responses that gave the information as it was literally stated in the studied text, and paraphrases that indicated gist comprehension of the studied information were scored as correct. Fact scores indicate the number of correctly answered questions about facts, ranging from 0 (none of the fact questions correct) to 5 (all fact questions correct) per text. Two independent raters scored all test responses of 60 participants (24% of the sample) on relation questions and of 58 participants (23%) for facts questions; inter-rater agreement was good for the number of correct relations (ICC = 0.86), and agreement was very high for the scoring of the fact questions (ICC = 0.99) (cf. Koo & Li, 2016). Therefore, coding was continued by one rater.

5.3.2. Scoring of diagrams

Three types of cues that could be derived from the filled-in text boxes in the completed and drawn diagrams were scored in line with Van Loon et al. (2014), who found these cues to be diagnostic (i.e., predictive of actual test performance): (1) correct response in the text box (a step in the causal chain is given; min = 0, max = 4), (2) commission error (incorrect response; min = 0, max = 4), (3) omissions (no response is given in this textbox, instead, the participant placed a question mark; min = 0, max = 4). In addition, we also coded the number of completed boxes, regardless of the accuracy of the content (min = 0, max = 4 for completed diagrams, max = 5 for drawn diagrams, as students in both diagram conditions were provided with one relation but this was a pre-filled box for the completion condition and a textual description of the relation in the instructions for the drawing condition; students in the drawing condition thus had to draw this relation themselves in their diagram). Because the range of scores differed across conditions, we converted the scores to percentages. Two raters independently coded diagram responses of 16 participants in the diagram completion condition (19% of the sample) and inter-rater agreement was good (ICC = 0.89). Independent scoring of responses in drawn diagrams of 16 participants (20% of the sample) showed high inter-rater agreement (ICC = 0.93).

5.4. Analyses

The measures used in this study are summarized in Table 1. All but one of these measures were gamma-correlations, a non-parametric correlation suitable to analyze ordinal JOL data (Nelson, 1984). Because diagram scores and test scores are measured on an interval scale, intra-individual Pearson correlations were used for analyses of cue diagnosticity.

For each gamma-measure, we only used those gammas that could be calculated based on at least four valid cases. Also, when a student's scores on a particular variable were invariant (e.g., all relation questions were answered correctly or all judgments about each text were equal) which renders calculation of the gamma correlation impossible, this student's score was omitted from the particular analysis. The missing values of the correlation-based measures ranged from 7.3% (for

Table 1
Summary of Measures Used in This Study.

Variable	Operationalization	Meaning
Monitoring accuracy	Gamma correlation between students' JOLs and the students' test performance (separately for facts and relations)	Values closer to +1 indicate high monitoring accuracy
Regulation accuracy	Gamma correlation between students' restudy selections (0 = not selected, 1 = selected) and the students' test score on relations questions	Values closer to −1 indicate high regulation accuracy
Relation monitoring and regulation	Gamma correlation between students' JOLs and their restudy selections (separately for facts and relations)	Values closer to + 1 indicate high consistency
Cue-diagnosticity	Pearson correlation between cues as coded in the diagrams (omissions, commission errors, number of correct relations, number of completed boxes) and a student's test performance (separately for facts and relations)	Number of correct relations and completed boxes: Values closer to + 1 indicate high diagnosticity Omissions and commission errors: Values closer to –1 indicate high diagnosticity
Cue-utilization	Gamma correlation between diagram cues (i.e., nr. of correct relations, omissions, commissions, and nr. of boxes completed) and students' JOLs of their test scores (separately for facts and relations)	Values closer to + 1 indicate high cue utilization

Note. JOL = judgment of learning

Table 2
Means and Standard Deviations per Condition for Students' Judgments of their Comprehension of Facts and Cause-and-effect Relations, Restudy Selections and Test Scores.

	Condition					
	Control		Diagram completion		Diagram drawing	
	M	(SD)	M	(SD)	M	(SD)
Judgments facts (0/20/40/60/80/100%)	54.92	(21.95)	48.61	(24.98)	54.26	(23.63)
Judgments relations (0/20/40/60/80/100%)	52.21	(23.23)	47.15	(26.66)	50.62	(25.23)
Restudy selection %[a] (0 = no, 1 = yes)	0.36	(0.48)	0.39	(49)	0.36	(0.48)
Test score fact questions (0–5)	1.25	(1.13)	1.16	(1.21)	1.10	(1.12)
Test score relation questions (0–4)	1.20	(1.14)	1.17	(1.19)	1.12	(1.12)

[a] This refers to the percentage of texts that were selected for restudy by students.

diagnosticity of correct relations) to 16.1% (for monitoring accuracy of facts).

Although our data had a nested structure (students were nested within teachers), we did not have enough level 2 units to conduct a multilevel analysis (Maas & Hox, 2005). To take the nested structure of the data into account, we included teacher dummies (i.e., 17 binary variables indicating each teacher) as covariates in our analyses (cf. McNeish & Stapleton, 2016).

We compared students' monitoring and regulation accuracy for comprehension judgments about relations and facts between the control condition and the diagram conditions (H1.1: monitoring and H1.2: regulation) and between the two diagram conditions. We conducted two MANCOVAS with planned contrasts; one MANCOVA on monitoring accuracy for facts and for relations (H1.1) and one on regulation accuracy for facts and for relations (H1.2), with condition as between-subjects factor and teacher dummies as covariates. The contrasts compared the control condition to both diagram conditions (H1.1/H1.2) and the two diagram conditions to each other.

To explore differences in cue-diagnosticity and cue-utilization between the two diagram conditions, we conducted MANCOVAs on the gamma correlations between the cue values (i.e., number of completed boxes, omissions, commission errors and correct relations) and students' test performance on facts and relations and between the cue values and students' JOLs for facts and relations, with condition (completion or drawing) as between-subjects factor and teacher dummies as covariates. Only in one instance, a teacher dummy yielded a significant effect.

5.5. Results

Table 2 shows mean JOLs, mean restudy selections, and mean test scores for relations and facts for the three conditions. A linear mixed model analysis[4] (level1 = text, level2 = student) showed that there were no differences between conditions in JOL magnitudes, the percentage of restudy selections, or students' test scores (JOLs facts: coeff. = −0.41, SE = 1.37, $p = .778$, $R^2 = 0.00$; JOLs relations: coeff. = −0.86, SE = 1.47, $p = .551$, $R^2 = 0.002$; restudy selections: coeff. = 0.01, SE = 0.06, $p = .907$, $R^2 = 0.334$; score facts: coeff. = −0.04, SE = 0.05, $p = .486$, $R^2 = 0.003$; score relations: coeff. = −0.08, SE = 0.05, $p = .163$, $R^2 = 0.01$).

5.5.1. Monitoring accuracy

Table 3 shows the gamma correlations indicating monitoring and regulation accuracy. In line with H1.1, contrasts revealed that students' monitoring accuracy for relations was significantly higher in the diagramming conditions than in the no-diagramming control condition, $F(1, 174) = 9.74$, $p = .002$, $\eta_p^2 = 0.053$, whereas there was no significant effect on students' monitoring accuracy of facts, $F(1, 174) = 0.25$, $p = .619$, $\eta_p^2 = 0.001$ (Also see Fig. 3A). There were no significant differences between the diagram completion condition and the diagram drawing condition in monitoring accuracy of relations, $F(1, 174) = 1.72$, $p = .191$, $\eta_p^2 = 0.010$, or facts, $F(1, 174) = 1.02$, $p = .314$, $\eta_p^2 = 0.006$.

5.5.2. Regulation accuracy

The gamma correlations between restudy selections and test performance for relations and fact questions are shown in Table 3. Contrary to H1.2, contrasts revealed no significant differences between the diagramming conditions and the no-diagramming control condition regarding students' regulation accuracy of relations, $F(1, 195) = 2.51$, $p = .114$, $\eta_p^2 = 0.013$, or facts, $F(1, 195) = 0.20$, $p = .654$, $\eta_p^2 = 0.001$.

There were no significant differences between the diagram completion condition and the diagram drawing condition in regulation accuracy for relations, $F(1, 195) = 0.675$, $p = .412$, $\eta_p^2 = 0.003$, or facts, $F(1, 195) = 2.05$, $p = .154$, $\eta_p^2 = 0.010$. To explore to what extent students used their JOLs to base their restudy decisions upon, we calculated the correlations between students' JOLs and restudy selections (Table 3). In all three conditions, students showed a high degree of consistency between their judgments for relations and facts and their restudy selections: all mean gamma correlations were significantly higher than zero (all $ps < 0.05$), suggesting that the restudy selections were strongly based on JOLs for the relations and facts. Differences in consistency between conditions were not significant (for relations: $F(2, 175) = 0.36$, $p = .70$, $\eta_p^2 = 0.004$; for facts: $F(2, 175) = 1.44$, $p = .24$, $\eta_p^2 = 0.016$).

5.5.3. Cue-diagnosticity and cue-utilization

Correlations between diagram responses and test performance (indicating cue diagnosticity) are shown in Table 4. The number of correct relations, omissions present in the diagrams and the number of boxes completed were significantly correlated to students' test scores on cause-and-effect relations. That is, these diagram cues were highly predictive of test performance on relations, both in the diagram completion and the diagram drawing condition. The number of commission errors was only diagnostic in the diagram completion condition.

Regarding fact test performance, only the number of correct relations showed a significant correlation with (i.e., was diagnostic of) students' facts scores for the drawing condition. All other cues (the number of commission errors, the number of boxes completed, and the number of omissions) were not significantly correlated (i.e., not diagnostic) to students' test scores on facts in either condition. There were no significant differences between the completion and the drawing condition in the correlations (i.e., the diagnosticity) of any of the cues (all $ps > 0.079$).

Table 4 also shows the correlation between diagram responses and JOLs (indicating cue utilization). Both in the diagram completion and in the diagram drawing condition, the number of completed boxes, correct relations, and omissions correlated with students' relation and fact JOLs, suggesting that they used these as cues for their JOLs. Regarding

[4] Linear mixed model analysis was used here because this analysis involved variables on the text level (as opposed to the analyses of the main hypotheses, that only involved variables on the student level).

Table 3
Means and Standard Deviations per Condition for Students' Monitoring Accuracy, Regulation Accuracy and the Relation Between JOLs and Restudy Selections (Gamma Correlations) and p-values for the Contrasts Tested.

	Control condition		Diagram completion condition		Diagram drawing condition		Contrast 1 (control vs diagrams)	Contrast 2 (diagram completion vs diagram drawing)
	M	(SD)	M	(SD)	M	(SD)	*p*	*p*
Monitoring accuracy facts	0.09	(0.60)	−0.01	(0.61)	0.08	(0.61)	0.62	0.31
Monitoring accuracy relations	0.00	(0.67)	0.43	(0.58)	0.28	(0.66)	0.002	0.19
Regulation accuracy facts	−0.10	(0.68)	0.003	(0.79)	−0.16	(0.75)	0.65	0.15
Regulation accuracy relations	−0.17	(0.72)	−0.31	(0.75)	−0.42	(0.68)	0.11	0.41
Relation between JOLs facts and restudy selections	−0.69	(0.58)	−0.73	(0.48)	−0.84	(0.39)	0.40	0.95
Relation between JOLs relations and restudy selections	−0.68	(0.59)	−0.76	(0.42)	−0.74	(0.41)	0.25	0.20

Table 4
Cue-Diagnosticity (Pearson Correlations) and Cue-Utilization (Gamma Correlations) for the two Diagram Conditions.

	Cue-diagnosticity relations (SD)	Cue-diagnosticity facts (SD)	Cue-utilization JOLs relations (SD)	Cue-utilization JOLs facts (SD)
Commission errors				
Completion	−.14[a] (0.45)	−0.04 (0.45)	0.10 (0.60)	0.12 (0.59)
Drawing	−0.05 (0.46)	−0.03 (0.52)	0.12 (0.58)	0.19 (0.62)
Nr of boxes completed				
Completion	.26[a] (0.42)	0.05 (0.42)	.71[a] (0.39)*	.60[a] (0.53)
Drawing	.28[a] (0.41)	0.08 (0.42)	.44[a] (0.56)*	.49[a] (0.59)
Nr of correct relations				
Completion	.39[a] (0.37)	0.07 (0.44)	.57[a] (0.45)*	.47[a] (0.52)
Drawing	.36[a] (0.43)	.12[a] (0.43)	.26[a] (0.60)*	.34[a] (0.64)
Omissions				
Completion	−.26[a] (0.42)	−0.05 (0.41)	−.71[a] (0.39)*	−.59[a] (0.52)
Drawing	−.28[a] (0.42)	−0.08 (0.42)	−.45[a] (0.56)*	−.49[a] (0.59)

* Means of diagram conditions are significantly different.
[a] Correlation differs significantly from 0, $p < .01$.

JOLs for facts, there were no significant differences in cue utilization between the two diagram groups (all *p*s > 0.05). Regarding JOLs for relations, however, there were: the number of completed boxes $F(1, 106) = 7.72$, $p = .006$, $\eta_p^2 = 0.07$, correct relations $F(1, 106) = 8.05$, $p = .005$, $\eta_p^2 = 0.07$, and omissions $F(1, 106) = 7.38$, $p = .008$, $\eta_p^2 = 0.07$ in students' diagrams showed higher correlations to students' JOLs in the diagram completion condition than in the diagram drawing condition. There was no significant difference in the correlation between the number of commission errors in students' diagrams and their JOLs for relations between the two conditions.

5.6. Discussion

The aim of this first experiment was to investigate the effects of diagram completion and drawing on students' monitoring and regulation accuracy of their text comprehension when studying complex texts. Replicating prior findings on diagram completion (Van Loon et al., 2014) and drawing (Schleinschok et al., 2017), performing a diagram task after studying texts and before making JOLs, was found to improve students' monitoring accuracy (as hypothesized, H1.1). Although the means seemed to indicate that monitoring for causal relations was somewhat better after diagram *completion* than after diagram *drawing*, this difference was not statistically significant.

The explorative analyses of cue diagnosticity and cue-utilization gave further insight into reasons why the diagram tasks may be beneficial for students' monitoring of their comprehension of cause-and-effect relations in texts. Students' ability to generate relations correctly (i.e., number of completed boxes and number of correct relations in the diagram) was predictive of their later test performance on relations. Similarly, their inability to generate relations (i.e., number of blank diagram boxes –omissions) predicted that they would not be able to come up with those relations at the test. Presumably, the diagram task gave students insight into what they did and did not understand, thereby benefitting monitoring accuracy of relations (i.e., students seem to have utilized these diagnostic cues when making JOLs). In line with our expectations and findings by Van Loon et al. (2014), however, this did not improve monitoring accuracy of facts, for which the diagram task hardly gave any diagnostic cues.

Students' judgments were strongly related to their restudy selections, which suggests that students used their monitoring judgments to make decisions about which texts needed to be restudied (cf. Metcalfe & Finn, 2008). However, in contrast to our hypothesis (H1.2) we did not find benefits of the diagram tasks for students' regulation accuracy (and in contrast to H1.4a/b, there were no differences in regulation accuracy between diagram conditions). This may be due to the fact that, even though diagram tasks improved monitoring accuracy, and students used their monitoring judgments in selecting texts for restudy, monitoring accuracy was only low to moderate, even after diagramming. This may be one explanation for why students failed to benefit from the diagram task when regulating their further learning (cf. Thiede et al., 2017). Another possible but speculative explanation is that students may not have selected all the texts that they thought they should restudy, because they wanted to finish the session sooner (i.e., students were not aware that they did not actually have to restudy the selected texts).

In sum, although students' monitoring of their text comprehension did improve from performing diagram tasks, it was still relatively low, and regulation did not improve (i.e., students often failed to select the texts they did not yet understand for further study). Students may need additional support in monitoring and regulating their text comprehension, and teachers could be an important source of such support, provided that they can accurately monitor their students' text comprehension and select those tasks for them that they need to study

further. Teachers' monitoring and regulation accuracy and effects of diagram completion and diagram drawing tasks on their accuracy, was addressed in Experiment 2.

6. Experiment 2: Teachers' monitoring and regulation accuracy

6.1. Method

6.1.1. Participants

The eighteen teachers of the students from Experiment 1 (56.56% women; age: $M = 37.00$, $SD = 11.30$; all were Dutch) participated in this study. They had on average 10.69 years of experience ($SD = 8.61$) in teaching their subject and had known their class on average for 12.17 months ($SD = 7.66$). They taught languages ($n = 10$), History ($n = 4$), Geography ($n = 2$), Biology ($n = 1$), or Economics ($n = 1$). We complied with the APA ethical standards for treatment of human participants, informed consent, and data management.

6.1.2. Research design

All teachers judged their students' text comprehension and made restudy selections for 15 randomly selected focus students who participated in Experiment 1. A within-subjects design was used; each teacher made judgments and restudy selections for five students in each condition (the control condition, the diagram completion condition, and the diagram drawing condition).

6.1.3. Materials

Similar to students in Experiment 1, teachers also received the instruction booklet (booklet 1) and the text reading booklet (booklet 2). Per focus student, the teacher received a JOL booklet (booklet 4) in which they had to indicate, per text, the percentage of questions about causal relations and of questions about facts each student would answer correctly at the test, and a restudy booklet (booklet 5) in which they indicated which text (s) each student should read again before taking the test (the order of the texts in these booklets matched the order that the focus student had). For making JOLs about focus students in the diagram completion and diagram drawing conditions, teachers received the diagram booklets that these students completed in Experiment 1 (i.e., booklet 3).

6.1.4. Procedure

All teachers participated in the experiment individually at their school. First, the experimenter and the teacher together read the practice booklet of the students so that the teacher was familiar with the (practice) tasks the students completed. Then, the teacher filled out a short questionnaire asking for demographics (e.g., gender, age, et cetera). The teachers received the text booklet and were given the opportunity to read the same six texts as the students had read at their own pace. Thereafter, the teachers practiced the judgment task by making judgments and restudy selections for three students, one per condition (random order); these judgments were not included in the analyses. Then, they made JOLs and afterwards, on a different page, teachers made the restudy selections for five students per condition (15 in total). The students of the three conditions were presented to the teacher in a random order. Teachers always saw the name of the student they were judging and were instructed to view the student's completed or drawn diagram, if available, prior to making JOLs for the students. Subsequently, the teacher completed the restudy selections for the same student. During the experiment, teachers were not allowed to look back and never received feedback. The total duration of the experiment was approximately 60 min.

6.1.5. Analyses

We used the same measures as in Experiment 1, but now linked the teachers' JOLs to the students' performance (i.e., monitoring accuracy), the teachers' restudy selection to the students' performance (i.e., regulation accuracy), the teachers' JOLs to their restudy selection, and the teachers' JOLs to the cues occurring in the diagrams (cue-utilization). Given that Experiment 2 had a within-subjects design, we conducted repeated-measures ANOVAs with two planned contrasts (cf. Experiment 1) to test our main hypotheses, one on monitoring accuracy and one on regulation accuracy of facts and relations, with condition as within subject factor. As in Experiment 1, we only used those gammas that were based on four or more valid cases and that had no invariance on either the teacher JOLs or the students' test scores. Percentages of missing values on the gamma-measures ranged from 5.6% to 11.6%. To explore differences in cue-utilization between the two diagram conditions, we conducted a repeated-measures ANOVA on the gamma correlations between the cue values (i.e., number of completed boxes, omissions, commission errors and correct relations) and teachers' JOLs for facts and relations, with condition (completion or drawing) as within subject factor.

Furthermore, we exploratively and descriptively compared students' and teachers' monitoring and regulation accuracy. Given the difference in the experimental designs (the student experiment having a between-subjects design and the teacher experiment a within-subjects design), we did not perform any statistical tests. Instead, we provide an interpretation of the differences in the means of students' and teachers' monitoring and regulation.

6.2. Results

Table 5 shows teachers' JOLs and restudy selections for their students. Linear mixed model analyses (level1 = text, level2 = student) showed that JOL magnitudes and the percentage of restudy selections did not differ between conditions (judgments facts: coeff. = 0.16, SE = 1.69, $p = .923$; judgments relations: coeff. = −0.00, SE = 1.52, $p = .998$; restudy selections: coeff. = −0.00, SE = 0.09, $p = .995$).

6.2.1. Monitoring accuracy

Table 6 shows the gamma correlations indicating teachers' monitoring and regulation accuracy. Contrary to what we expected (H2.1), there were no significant differences between teachers' monitoring accuracy when viewing students' diagrams compared to not having diagrams available, either for relations, $F(1, 15) = 2.08$, $p = .170$, $\eta_p^2 = 0.122$, or facts, $F(1, 15) = 2.61$, $p = .127$, $\eta_p^2 = 0.148$. In addition, there were no significant differences in teachers' monitoring accuracy when viewing completed diagrams compared to viewing drawn diagrams, either for relations, $F(1, 15) = 0.002$, $p = .969$, $\eta_p^2 = 0.000$, or facts, $F(1, 15) = 0.08$, $p = .782$, $\eta_p^2 = 0.005$.

Table 5
Means and Standard Deviations per Condition for Teachers' Judgments of Students' Comprehension of Facts and Cause-and-effect Relations and Teachers' Restudy Selections.

	Condition					
	Control		Diagram completion		Diagram drawing	
	M	(SD)	M	(SD)	M	(SD)
Judgments facts (0/20/40/60/80/100%)	54.62	(24.48)	54.55	(23.46)	57.06	(23.76)
Judgments relations (0/20/40/60/80/100%)	59.78	(23.40)	58.33	(24.28)	60.31	(23.66)
Restudy selection (0 = no, 1 = yes)	0.27	(0.44)	0.30	(0.46)	0.27	(0.44)

Table 6
Means and Standard Deviations per Condition for Teachers' Monitoring, Regulation, and Relation Between JOLs and Restudy Choices (Gamma Correlations) and p-values for the Contrasts Tested.

	Control condition		Diagram completion condition		Diagram drawing condition		Contrast 1 (control vs diagrams)	Contrast 2 (diagram completion vs diagram drawing)
	M	(SD)	M	(SD)	M	(SD)	*p*	*p*
Monitoring accuracy facts	−0.09	(0.39)	0.08	(0.36)	0.12	(0.38)	0.13	0.78
Monitoring accuracy relations	0.03	(0.58)	0.26	(0.36)	0.26	(0.34)	0.17	0.97
Regulation accuracy facts	0.08	(0.47)	−0.18	(0.44)	0.04	(0.50)	0.22	0.29
Regulation accuracy relations	−0.00	(0.52)	−0.38	(0.37)	−0.21	(0.51)	0.03	0.27
Relation JOL facts – restudy	−0.73	(0.29)	−0.71	(0.42)	−0.89	(0.21)	0.53	0.08
Relation JOL relations - restudy	−0.77	(0.24)	−0.81	(0.23)	−0.87	(0.19)	0.38	0.50

6.2.2. Regulation accuracy

The gamma correlations between teachers' restudy selections and test performance (i.e., regulation accuracy) for relation and fact questions are shown in Table 6. In line with our hypothesis (H2.2), teachers' regulation accuracy when viewing students' diagrams (i.e., completed or drawn) was significantly higher than when not viewing students' diagrams for relations, $F(1, 16) = 5.53$, $p = .032$, $\eta_p^2 = 0.26$, but not for facts $F(1, 16) = 1.65$, $p = .217$, $\eta_p^2 = 0.09$.

There were no significant differences in teachers' regulation accuracy between the diagram completion condition and the diagram drawing condition regarding relations, $F(1, 16) = 1.31$, $p = .270$, $\eta_p^2 = 0.08$, or facts, $F(1, 16) = 1.18$, $p = .294$, $\eta_p^2 = 0.07$. To explore to what extent teachers used their JOLs to base their restudy decisions upon, we calculated the correlations between teachers' JOLs and restudy selections (Table 6). These correlations indicate that in all three conditions, teachers showed a high degree of consistency between their judgments for relations and facts and their restudy selections (all mean gamma correlations were significantly higher than zero, all $ps < 0.05$).

6.2.3. Exploring teachers' Cue-Utilization

Table 7 shows the relation between students' diagram responses and teachers' JOLs (cue utilization). Both in the diagram completion and the diagram drawing conditions, teachers used the number of completed boxes, correct relations, and omissions as cues for their relation and fact JOLs. As shown in Table 4, these cues were indeed diagnostic of students' comprehension of causal relations, but not of facts (with the exception of the number of correct relations in the diagram drawing condition, which did correlate significantly with performance on facts questions). Furthermore, Table 7 shows that when viewing *drawn* diagrams, teachers also used the number of commission errors when judging students' comprehension of relations and facts and made use of this cue to a greater extent than when teachers viewed completed diagrams both when monitoring students' comprehension of relations, $F(1,17) = 9.59$, $p = .007$, $\eta_p^2 = 0.36$, and of facts, $F(1,17) = 6.62$, $p = .020$, $\eta_p^2 = 0.28$. Yet this cue was only diagnostic for students' comprehension of relations in the *completion* condition (Table 4). There were no significant differences between conditions for the other cues.

Table 7
Teachers' Cue-Utilization (Gamma Correlations) for Making Judgments of Students' Comprehension of Cause-and-Effect Relations.

	Cue-utilization JOLs relations (SD)	Cue-utilization JOLs facts (SD)
Commission errors		
Completion	−0.04 (0.41)*	−0.07 (0.34)*
Drawing	.35[a] (0.35)*	.22[a] (0.36)*
Nr of boxes completed		
Completion	.68[a] (0.30)	.60[a] (0.34)
Drawing	.57[a] (0.33)	.61[a] (0.30)
Nr of correct relations		
Completion	.54[a] (0.23)	.49[a] (0.26)
Drawing	.39[a] (0.36)	.49[a] (0.36)
Omissions		
Completion	−.67[a] (0.30)	−.60[a] (0.34)
Drawing	−.56[a] (0.35)	−.61[a] (0.30)

* Means of diagram conditions are significantly different.
[a] Correlation differs significantly from 0, $p < .01$.

7. Comparison between students' and teachers' monitoring and regulation accuracy

7.1. Monitoring accuracy

Fig. 2 shows students' and teachers' monitoring accuracy. When comparing students' and teachers' monitoring, we see that their accuracy is quite similar in the different conditions. Although statistical comparison between student and teacher gammas is not possible due to due to differences in the design of Experiments 1 (between-subjects) and 2 (within-subjects), visual exploration of differences between students' and teachers' mean gamma correlations suggests that in the diagram completion condition, monitoring of relations is possibly somewhat more accurate for students than for teachers whereas monitoring of facts is possibly somewhat more accurate for teachers than for students.

7.2. Regulation accuracy

In the majority of the cases, a teacher and a student selected the same texts (63%). In 23% of the cases, a student selected a text that was not selected by the teacher. In 14% of the cases, a teacher selected a text that was not selected by the student. Furthermore, teachers appeared to select less texts (M = 2.5) than students (M = 3).When visually inspecting the means, we see that (1) in the control condition, regulation of facts and relations is possibly somewhat more accurate for students than for teachers, (2) in the diagram completion condition, regulation accuracy of facts (not of relations) is possibly somewhat more accurate for teachers than for students, and (3) in the diagram drawing condition, regulation of facts and relations is possibly somewhat more accurate for students than for teachers. Yet, again, this conclusion is only based on inspection of the means and we cannot conclude that these visual differences are statistically significant.

7.3. Discussion

In Experiment 2, teachers judged their students' comprehension of causal relations and facts from texts, either without diagrams or seeing

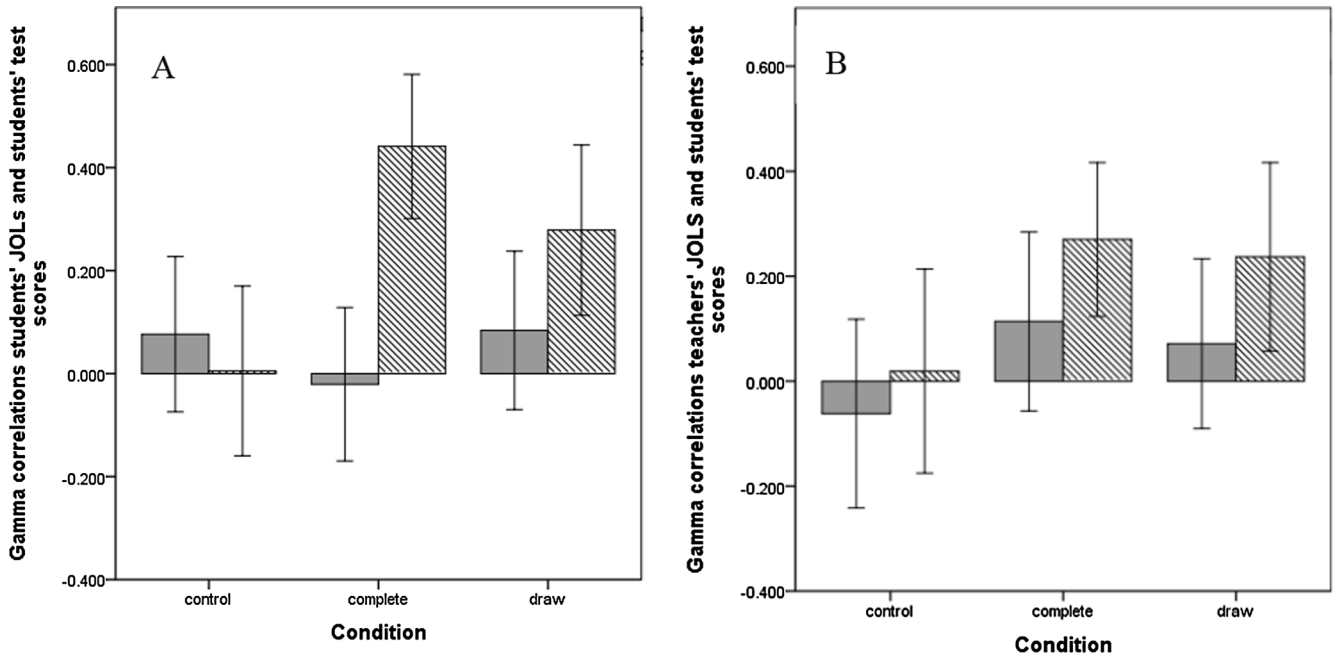

Fig. 2. Students' (A) and Teachers' (B) monitoring accuracy of facts (grey) and relations (shaded). Note. Error bars represent 95% confidence intervals.

students' completed or hand-drawn diagrams. After making those monitoring judgments, they selected texts that students should restudy (i.e., regulation). Findings show that, in line with our hypothesis (H2.2) when having access to representations of their students' text comprehension (i.e., diagrams), the regulation accuracy of these teachers increased considerably with regard to cause-and-effect relations (i.e., teachers selected those relations for restudy that students indeed did not understand), but not with regard to students' factual understanding. Because monitoring accuracy is often considered a necessary but not sufficient condition for regulation accuracy, it is surprising that regulation accuracy improved from seeing diagrams while the accuracy of teachers' monitoring of their students' comprehension of relations did not (in contrast to our hypothesis H2.1). These findings will be discussed in the next section.

8. General discussion

The aim of the present study was to investigate whether two generative activities (i.e., diagram completion and diagram drawing) performed by students after text study, would improve the accuracy of students' and teachers' monitoring judgments of students' text comprehension and the accuracy of subsequent regulation of study activities (i.e., deciding which texts should be restudied). Diagrams (whether completed or drawn) were helpful for both students and teachers, yet for different types of judgments: Diagramming improved students' *monitoring* accuracy, that is, their ability to judge which texts they understood better than other texts, whereas viewing students' diagrams (i.e., representations of their students' text comprehension) improved teachers' *regulation* accuracy.

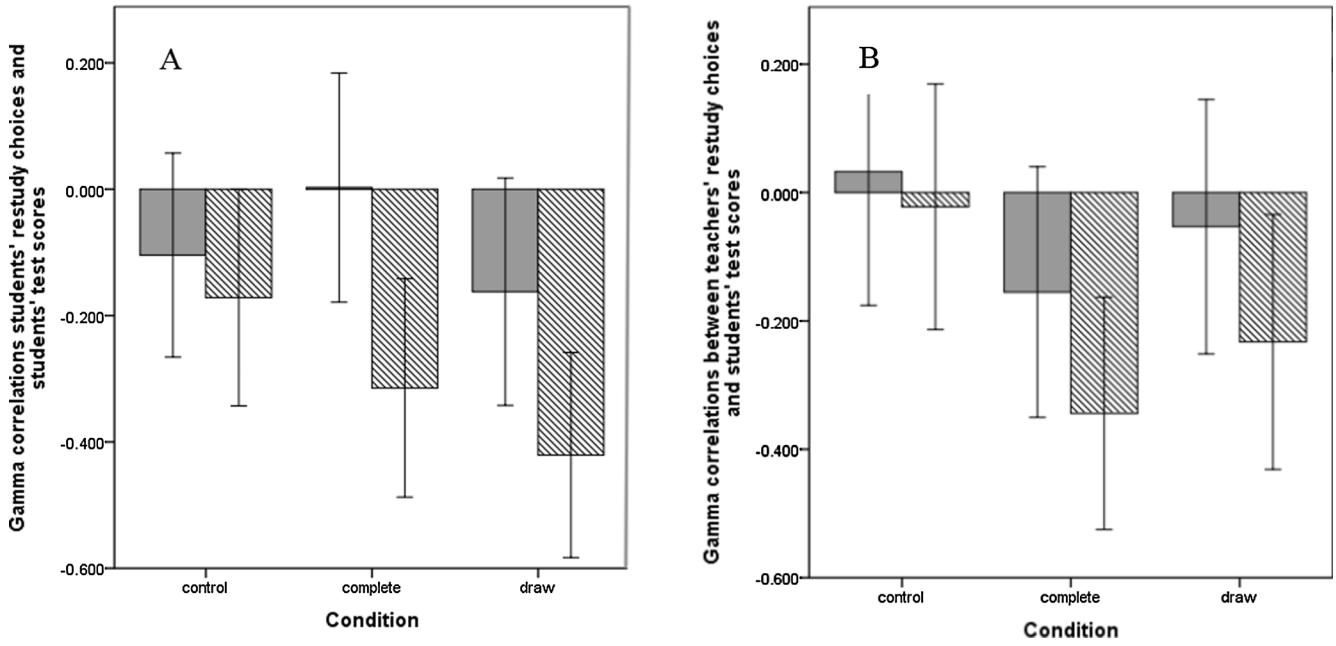

Fig. 3. Students' (A) and Teachers' (B) regulation accuracy of facts (grey) and relations (shaded). Note. Error bars represent 95% confidence intervals.

Our findings regarding students' monitoring corroborate and extend findings by Van Loon et al. (2014) who investigated effects of (delayed) diagram completion and similarly found that it improved the accuracy of students' monitoring of their text comprehension. Our results further suggest (in line with Van Loon et al.) that this effect presumably arises because diagramming generates cues that are diagnostic (i.e., predictive) of students' comprehension of causal relations (i.e., omissions, commission errors, the number of correct relations and – additionally included in the current study – the number of completed boxes). These cues were diagnostic both in completed and hand-drawn diagrams (exception: commission errors were only diagnostic in the completion condition) and our analyses show that students actually used several of those cues while making monitoring judgments (i.e., omissions, commission errors, the number of correct relations, and number of boxes completed).

The results from Experiment 1 underline the effectiveness of generative activities for improving students' monitoring accuracy, and suggests that it does not matter whether students complete partially pre-defined diagrams created by teachers or instructional designers (cf. Van Loon et al., 2014, with texts; see also Baars et al., 2013, with worked examples) or have to draw them from scratch. Our study also adds to findings regarding the effectiveness of drawing for monitoring accuracy, suggesting that next to drawing the conceptual content of a text (e.g., Kostons & de Koning, 2017; Schleinschok et al., 2017), drawing the causal relations contained in a text may also be a fruitful and easy to implement way to improve monitoring accuracy.

Interestingly, our findings from Experiment 2 show that having access to the products of students' generative activities also benefits *teachers*. Surprisingly, however, given that monitoring accuracy is often considered a necessary but not sufficient condition for regulation accuracy, seeing students' completed or drawn diagrams only improved teachers' regulation accuracy (i.e., their ability to determine which texts a student should restudy), not their monitoring accuracy.

The finding that having access to students' diagrams improved regulation accuracy corroborates findings of Van de Pol et al. (2014) showing that teachers' regulation fitted students' comprehension better (i.e., regulation accuracy) when they gathered more diagnostic information about students' comprehension. These findings seem to imply that teachers can make more adaptive restudy decisions for individual students when they have access to diagnostic cues about students' comprehension. It is promising for educational practice that this effect was obtained with a relatively simple and easy to implement intervention in our study. Even though further research would be needed before recommending implementation in practice, this technique would be relatively easy to use and implement, and could easily be taught to teachers in initial teacher education or continued professional development courses.

Visual and explorative inspection suggests that students' and teachers' mean monitoring and regulation accuracy was quite similar. One potential difference occurred with regard to insight into comprehension of cause-and-effect relations in the text. Students who completed diagrams possibly monitored their text comprehension somewhat more accurately than teachers who observed these completed diagrams. Furthermore, students who drew diagrams may possibly have made more accurate regulation decisions than teachers who observed these drawn diagrams. Experiential information (see Koriat, 1997) that students acquire in the process of completing or drawing diagrams, to which teachers have no access, may provide them with important cues about their text comprehension. These findings should be interpreted with caution though, as confident intervals are wide and given that we did not test for statistical significance due to differences in the design of Experiments 1 (between-subjects) and 2 (within-subjects). Yet, this tentative finding may be useful as a starting point for future research.

In general, even though generating or observing diagrams affected accuracy, it should be noted that both students' and teachers' overall levels of monitoring and regulation accuracy were still relatively low. A possible explanation for this relatively low monitoring and regulation accuracy could be that, even though students and teachers appeared to use the diagnostic cues with which the diagrams provided them, such as the number of correct relations or omissions, they might still have usednon-diagnostic cues which they also had to their disposal. Students, for example, might have used experiential cues that were not necessarily predictive of their comprehension such as their feelings of processing fluency (Mueller, Tauber, & Dunlosky, 2013) or their beliefs about their domain knowledge of the studied texts (Griffin, Jee, & Wiley, 2009). Teachers might have used their general knowledge about the student, which may not always be diagnostic for their actual performance. For instance, Kaiser et al. (2015), using vignettes of fictional students, showed that teachers' monitoring accuracy of students' math performance decreased when they had both diagnostic (e.g., oral/written mathematics proficiency) and non-diagnostic (e.g., family background) information available about the student, compared to having only diagnostic information available. Oudman, Van de Pol, Bakker, Moerbeek, and Van Gog (2018), studying primary teachers' judgments of students' mathematical performance, showed similar results. That is, teachers' judgments were more accurate when they only had access to diagnostic cues (i.e., work of an anonymous student) than when they also had access to non-diagnostic cues (i.e., student cues).

This brings us to a limitation of the present study: we did not directly measure students' and teachers' cue-utilization. Although we were able to establish (cf. Van Loon et al., 2014) which cues obtained from students' diagrams were diagnostic of (i.e., correlated with) students' performance and whether these correlated with students' and teachers' judgments, we do not know exactly which cues they considered at the time of making monitoring and regulation judgments and we have no information about cue-utilization in the control condition. Future research should address this question of what cues students and teachers use, for instance by using concurrent or retrospective verbal reports (cf. Oudman et al., 2018). Teachers' use of student-related cues and the extent to which these are diagnostic could be experimentally investigated by manipulating the availability of cues. For instance, by including an extra condition in which teachers do not know from which student a diagram originates while making judgments, one could infer the effect of student-related cues (which were available in the no diagram control condition and the diagram conditions in the present study).

Another potential limitation of this study is that, even though students and teachers received information about the test and saw some examples of the type of test questions in the instruction session, they did not know the exact content of the test. That this might have affected their judgment accuracy is suggested by research showing that students' JOL accuracy may improve when they know the content of the test questions, and when they attempt answering these questions prior to making judgments (Dunlosky, Rawson, & Middleton, 2005). Similarly, Südkamp et al. (2012) showed that teachers' monitoring accuracy increased when they "were informed about the achievement test on which their judgment of student achievement would be based" (p. 749). In addition, teachers were not given the correct answers to the test questions and not all teachers may have been knowledgeable about all text topics. Ostermann, Leuders, and Nückles (2017) showed that an intervention aimed at increasing teachers' pedagogical content knowledge (i.e., knowledge about task characteristics and students' misconceptions), improved teachers' monitoring accuracy, as teachers were presumably better able to assume the student's perspective. Future research should therefore investigate whether informing students and teachers about the exact test questions would further increase monitoring and regulation accuracy, and whether providing teachers with the correct test answers, giving them background information about the topics and students' misconceptions, would promote their monitoring and/or regulation accuracy. In addition, although the N at the student level was high (N = 248), the N at the teacher level was relatively low (N = 18). Given the size of this sample, the results of this study should be interpreted cautiously.

In addition, the design of this study does not allow us to determine whether increased monitoring/regulation accuracy accomplished with the use of diagrams subsequently increased students' achievement. However, previous studies (e.g., Kornell & Metcalfe, 2006; Metcalfe &

Finn, 2013; Thiede et al., 2003) showed that, when allowing for restudy, increased monitoring/regulation accuracy improved students' achievement.

Furthermore, we assessed comprehension by means relation questions; however, it would be interesting to include a wider array of reading comprehension measures in future research; for instance, a measure that goes even further in assessing students' deep understanding would be to ask them to apply the information from the text in a problem-solving task (cf. McNamara & Kendeou, 2017).

Finally, a potential limitation induced by our experimental setting is that it may have inhibited students' existing text comprehension strategies. However, it is known that the strategies that students prefer, are not necessarily the ones that are most effective for their learning (e.g., Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013), so it is unclear to what extent inhibiting their existing strategies (and replacing them with other strategies such as diagramming) would help or hinder students' monitoring and regulation accuracy. This would be an interesting issue for future research to address.

9. Conclusion

To conclude, our study showed, in line with prior research, that the key to improving students' monitoring accuracy when learning from texts lies in providing them with access to cues that are diagnostic of their text comprehension. Both having students complete or draw diagrams seems an effective way to do so, although there is still room for further improvement in monitoring accuracy, and it remains an open question how to improve their subsequent regulation accuracy. A novel contribution of our study further lies in showing that it is also beneficial for teachers to provide them with access to cues that are diagnostic of students' text comprehension. Seeing the product of students' diagrams (and possibly other generative activities) improved teachers' regulation accuracy, which is important for making adaptive instructional decisions and ultimately for students' academic achievement.

Acknowledgments

We would like to thank all students who helped collecting the data, and Kirsten van Pelt and Anniek de Kort for their help in coding the students' tests and diagrams. During the realization of part of this research the first author was funded by a Veni grant from the Netherlands Organization for Scientific Research awarded to the first author (grant number: 451-16-012).

Appendix A. Example Texts

A.1. Text "The Suez Canal"

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

A.2. Text "Botox"

Botox is the abbreviation of BotuliniumToxin, this is a poison that is produced by the bacterium Clostridiumbotulinum. This substance blocks the signal between the nerves and the muscles in the skin. Since 1989, use of Botox is permitted, although this is strictly controlled in The Netherlands. In 2004, 28 people died in America, they had an accident with an incorrect dosage of Botox. Due to the blocking of the signal between the nerves and skin, originally, Botox was particularly used against muscle contractions, for example with patients who could not control muscle contractions and continuously blinked their eyes. By injecting Botox around the eyes, the muscles are paralyzed and the muscle contractions disappear. Because Botox blocks the signal between the nerves and the muscles in the skin, this is also used in plastic surgery to smoothen the skin: It can reduce the wrinkles around the eyes and the forehead. Because wrinkles are reduced, this treatment makes people look younger. The effect of such a treatment usually lasts between 1 and 6 months. However, this treatment against wrinkles between the eyes and on the forehead can also undesirably change peoples' face expressions.

Reprinted with permission from "Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy," by M. Van Loon et al., 2014, Acta Psychologica, 151, p. 145. Copyright 2014 by M van Loon.

Appendix B. Example Test Questions

B.1. Questions about causal relations

Text "Suez"

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

Text "Botox"

Botox blocks the signal between the nerves and the skin. What are the effects of this?

Questions about factual information

Text "Suez"

- In what year was the Suez Canal opened for ships?
- From which country was the engineer who designed the Suez Canal?

Text "Botox"

- What is the full name of Botox?
- Since when has use of Botox been officially permitted?

Reprinted from "Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy," by Van Loon et al. (2014), Acta Psychologica, 151, p. 145. Copyright 2014 by M van Loon. Reprinted with permission.

Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cedpsych.2019.02.001.

References

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

Baars, M., van Gog, T., de Bruin, A., & Paas, F. (2017). Effects of problem solving after worked example study on secondary school children's monitoring accuracy. *Educational Psychology, 37*(7), 810–834. https://doi.org/10.1080/01443410.2016.1150419.

Behrmann, L., & Souvignier, E. (2015). Effects of fit between teachers' instructional beliefs and didactical principles of reading programs. *European Journal of Psychology of Education, 30*(3), 295–312. https://doi.org/10.1007/s10212-014-0241-6.

Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological Bulletin, 52*, 177–193. https://doi.org/10.1037/h0044919.

De Bruin, A. B., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology, 109*(3), 294–310. https://doi.org/10.1016/j.jecp.2011.02.005.

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

Dunlosky, J., Mueller, M. L., & Tauber, S. K. (2014). The contribution of processing fluency (and beliefs) to people's judgments of learning. In D. Stephen Lindsay, Colleen M. Kelley, Andrew P. Yonelinas, Henry L. Roediger III (Eds), Remembering: Attributions, processes, and control in human memory: Papers in honour of Larry L. Jacoby (pp. 46–64). Hove, UK: Psychology Press.

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

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4–58. https://doi.org/10.1177/1529100612453266.

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language, 52*(4), 551–565. https://doi.org/10.1016/j.jml.2005.01.011.

Finn, B., & Tauber, S. K. (2015). When confidence is not a signal of knowing: How students' experiences and beliefs about processing fluency can lead to miscalibrated confidence. *Educational Psychology Review, 27*(4), 567–586. https://doi.org/10.1007/s10648-015-9313-7.

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

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

Herppich, S., Praetorius, A., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., & Südkamp, A. (2019). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education.*. https://doi.org/10.1016/j.tate.2017.12.001 in press.

Kaiser, J., Möller, J., Helm, F., & Kunter, M. (2015). Das schülerinventar: Welche Schülermerkmale die Leistungsurteile von Lehrkräften beeinflussen. *Zeitschrift Fur Erziehungswissenschaft, 18*(2), 279–302. https://doi.org/10.1007/s11618-015-0619-5.

Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. *Learning and Instruction, 28*, 73–84. https://doi.org/10.1016/j.learninstruc.2013.06.001.

Karing, C., Pfost, M., & Artelt, C. (2011). Hängt die diagnostische Kompetenz von Sekundarstufenlehrkräften mit der Entwicklung der Lesekompetenz und der mathematischen Kompetenz ihrer Schülerinnen und Schüler zusammen? Retrieved from *Journal for Educational Research Online, 3*(2), 121. <https://www.pedocs.de/volltexte/2012/5626/pdf/JERO_2011_2_Karing_Pfost_Artelt_Haengt_die_diagnostische_Kompetenz_D_A.pdf>.

Kimball, D. R., & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory & Cognition, 31*, 918–929. https://doi.org/10.3758/Bf03196445.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review, 95*, 163–182.

Klug, J., Bruder, S., Kelava, A., Spiel, C., & Schmitz, B. (2013). Diagnostic competence of teachers: A process model that accounts for diagnosing learning behavior tested by means of a case scenario. *Teaching and Teacher Education, 30*, 38–46. https://doi.org/10.1016/j.tate.2012.10.004.

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

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

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. https://doi.org/10.1037/0096-3445.126.4.349.

Kostons, D., & de Koning, B. B. (2017). Does visualization affect monitoring accuracy, restudy choice, and comprehension scores of students in primary education? *Contemporary Educational Psychology, 51*, 1–10. https://doi.org/10.1016/j.cedpsych.2017.05.001.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 86–92. https://doi.org/10.1027/1614-2241.1.3.86.

McNamara, D. S., & Kendeou, P. (2017). Translating advances in reading comprehension research to educational practice. *International Electronic Journal of Elementary Education, 4*(1), 33–46.

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28*(2), 295–314. https://doi.org/10.1007/s10648-014-9287-x.

Metcalfe, J., & Finn, B. (2013). Metacognition and control of study choice in children. *Metacognition and Learning, 8*(1), 19–46. https://doi.org/10.1007/s11409-013-9094-7.

Mihalca, L., Mengelkamp, C., & Schnotz, W. (2017). Accuracy of metacognitive judgments as a moderator of learner control effectiveness in problem-solving tasks. *Metacognition and Learning, 1–23*. https://doi.org/10.1007/s11409-017-9173-2.

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review, 20*(2), 378–384. https://doi.org/10.3758/s13423-012-0343-6.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*(1), 109. https://doi.org/10.1037/0033-2909.95.1.109.

Ostermann, A., Leuders, T., & Nückles, M. (2017). Improving the judgment of task difficulties: Prospective teachers' diagnostic competence in the area of functions and graphs. *Journal of Mathematics Teacher Education, 1–27*. https://doi.org/10.1007/s10857-017-9369-z.

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

Pino-Pasternak, D., Whitebread, D., & Tolmie, A. (2010). A multidimensional analysis of parent–child interactions during academic tasks and their relationships with children's self-regulated learning. *Cognition and Instruction, 28*(3), 219–272. https://doi.org/10.1080/07370008.2010.490494.

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.

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.

Südkamp, A., Möller, J., & Pohlmann, B. (2008). Der Simulierte Klassenraum: Eine experimentelle Untersuchung zur diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie, 22*, 261–276. https://doi.org/10.1024/1010-0652.22.34.261.

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

Thiede, K. W., & De Bruin, A. B. H. (2017). Self-regulated learning in reading. In D. Schunk, & J. Greene (Eds.). *Handbook of Self-Regulation of Learning and Performance* (pp. 24–137). New York: Routledge.

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

Thiede, K. W., Brendefur, J. L., Osguthorpe, R. D., Carney, M. B., Bremner, A., Strother, S., ... Jesse, D. (2015). Can teachers accurately predict student performance? *Teaching and Teacher Education, 49*, 36–44. https://doi.org/10.1016/j.tate.2015.01.012.

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1267. https://doi.org/10.1037/0278-7393.31.6.1267.

Thiede, K. W., Redford, J. S., Wiley, J., & Griffin, T. D. (2017). How restudy decisions affect overall comprehension for seventh-grade students. *British Journal of Educational Psychology*. https://doi.org/10.1111/bjep.12166.

Van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review, 22*(3), 271–296. https://doi.org/10.1007/s10648-010-9127-6.

Van de Pol, J., Volman, M., Oort, F., & Beishuizen, J. (2014). Teacher scaffolding in small-group work: An intervention study. *Journal of the Learning Sciences, 23*(4), 600–650. https://doi.org/10.1080/10508406.2013.805300.

Van de Pol, J., Volman, M., Oort, F., & Beishuizen, J. (2015). The effects of scaffolding in the classroom: Support contingency and student independent working time in relation to student achievement, task effort and appreciation of support. *Instructional Science, 43*(5), 615–641. https://doi.org/10.1007/s11251-015-9351-z.

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.

Vermunt, J. D., & Verloop, N. (1999). Congruence and friction between learning and teaching. *Learning and instruction, 9*(3), 257–280. https://doi.org/10.1016/S0959-4752(98)00028-0.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.). *Metacognition in educational theory and practice* (pp. 277–304). New York: Routledge.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x.