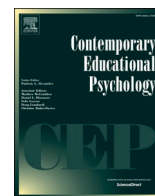




Contents lists available at ScienceDirect

Contemporary Educational Psychology

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

The role of feedback on students' diagramming: Effects on monitoring accuracy and text comprehension

Sophia Braumann^{a,*}, Janneke van de Pol^a, Ellen Kok^a, Héctor J. Pijeira-Díaz^b, Margot van Wermeskerken^a, Anique B.H. de Bruin^b, Tamara van Gog^a

^a Department of Education, Utrecht University, Heidelberglaan 1, P.O. Box 80140, 3508 TC Utrecht, the Netherlands

^b Department of Educational Development and Research, and School of Health Professions Education, Maastricht University, P.O. Box 616, 6200 MD Maastricht, the Netherlands

ARTICLE INFO

Keywords:

Monitoring accuracy
Text comprehension
Causal diagramming
Feedback standards
Cue-utilization
Eye-tracking
Cued retrospective report

ABSTRACT

Accurate self-monitoring of text comprehension is critical for effective self-regulated learning from texts. Unfortunately, it has been repeatedly shown that students' monitoring of their text comprehension is often inaccurate, which can subsequently lead to inaccurate regulation and ineffective restudy decisions. Previous research provided evidence that completing causal diagrams at a delay after text reading (i.e., diagramming) can help to improve students' monitoring of text comprehension. However, even after diagramming, there is still substantial room for improvement. The current studies therefore aimed to test whether providing feedback in the form of a correctly completed diagram (i.e., performance standard) would further increase students' monitoring accuracy. In Study 1, 79 participants (aged 18–23) made judgements of learning under four conditions: I. No-Diagram (control), II. Standard-Only, III. Diagramming-Only, or IV. Diagramming + Standard. In each condition, students studied a text, made a judgement of learning before and after the experimental tasks, and completed a comprehension test at the end of each of the (overall six) trials. Results showed that only Diagramming + Standard improved monitoring accuracy and text comprehension. In Study 2, 20 undergraduate students (aged 18–23) completed the Diagramming + Standard condition while their eye movements were tracked and subsequently replayed for cued retrospective verbal reporting. The findings suggest that students used the standards to identify mistakes and improve their monitoring and text comprehension.

1. Introduction

When studying texts, students need to self-regulate their learning to learn effectively. That is, they need to accurately monitor and regulate their learning process, for example by (partially) restudying texts that are not yet well understood (Nelson & Leonesio 1988; Thiede & Dunlosky 1999). Particularly, accurately judging one's text comprehension (also referred to as *metacomprehension*; Maki & Berry, 1984) is critical for self-regulated learning as it is often the basis for effective restudy decisions and study planning (Hacker & Bol, 2019). Unfortunately, students' monitoring judgements of their text comprehension are often inaccurate (Lipko & Dunlosky, 2007; Maki, 1998; Prinz et al., 2020a). Consequently, they make suboptimal (re)study decisions, and thereby limit their learning outcomes (Thiede & Dunlosky, 1999; Thiede et al., 2003; Winne & Hadwin, 2010). Thus, researchers have looked for interventions that aim to improve monitoring accuracy of text

comprehension and have successfully identified several generative activities that do so (cf. Prinz et al., 2020b). Among those generative activities for causal texts is diagramming (i.e., having students complete a diagram of causal relations in the text; van Loon et al., 2014). Completing causal diagrams is generally a particularly suitable generative task for improving monitoring accuracy, as it is relatively simple to implement in comparison to, for example, concept mapping which requires extensive training (Redford et al., 2012). However, there is room for improvement in monitoring accuracy even after such interventions and interventions that improve both monitoring accuracy and text comprehension are scarce (Hacker & Bol, 2019).

In two studies, we investigate whether offering students feedback in the form of correct diagrams that can serve as performance standards (hereafter *standards*) after diagramming, would further enhance monitoring accuracy and text comprehension. Comparing their own diagrams to standards could help students to assess the quality of their answers (e.

* Corresponding author at: Utrecht University, Heidelberglaan 1, P.O. Box 80140, 3508 TC Utrecht, the Netherlands.

E-mail address: s.e.braumann@uu.nl (S. Braumann).

<https://doi.org/10.1016/j.cedpsych.2023.102251>

Available online 14 November 2023

0361-476X/© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

g., identify mistakes) and thereby improve their monitoring judgements (Dunlosky et al., 2011; Rawson & Dunlosky, 2007; Waldeyer & Roelle, 2020). At the same time, students can learn from the standard itself and thereby improve their text comprehension (McCrudden et al., 2007).

However, for a standard to be effective for improving monitoring accuracy, students would need to systematically process it by comparing their own responses to the standard. It is currently unclear to what extent students are able to do so. Process measures such as eye tracking (Holmqvist et al., 2011) and verbal protocols (van Gog et al., 2005) can provide insights into how students process the standard to assess the quality of their own responses. Hence, the first aim of the present studies was to investigate the effect of providing a standard, in the form of a correctly completed diagram after a diagramming task, on students' monitoring accuracy and text comprehension (Study 1). The second aim was to explore how students processed the standard and what information they extracted from the comparison of their own and the standard diagram to inform their monitoring judgements of their text comprehension (Study 2).

1.1. Diagramming for improving monitoring accuracy

Many interventions that aim to improve students' monitoring accuracy of their text comprehension are inspired by Koriat's (1997) Cue Utilization Framework. The framework distinguishes between diagnostic cues (information predictive of actual comprehension test performance) and non-diagnostic cues (information not predictive of actual comprehension test performance) that students may use when judging their comprehension. Research has shown that non-diagnostic cues in this context are, for example, the text length or students' interest in the text (i.e., these were not predictive of comprehension; Jaeger & Wiley,

2014; Thiede et al., 2010; also in a study using the same materials: van de Pol et al., 2021). Diagnostic cues can be, for example, boxes left empty in a causal diagram completed after reading (i.e., these indicate gaps in comprehension; van de Pol et al., 2020). When students make use of non-diagnostic cues, their monitoring accuracy tends to be low (Thiede et al., 2010), so interventions to improve monitoring accuracy often aim to foster the use of diagnostic cues for making monitoring judgements (Prinz et al., 2020b).

Research has shown that performing *generative activities* (see Griffin et al. (2019) for an overview), which require students to generate information on the gist of the text, provide students with more predictive (or *diagnostic cues*) regarding how well they have understood a text. Examples of generative activities are listing keywords (De Bruin et al., 2011), writing summaries (Thiede & Anderson, 2003), drawing concept maps (Thiede et al., 2010), or completing causal diagrams (van Loon et al., 2014; van de Pol et al., 2019). In theory, monitoring judgements of their text comprehension would become more accurate if students subsequently also base their judgements on these diagnostic cues (Thiede et al., 2019). Having students complete causal diagrams (see Fig. 1 for an example) can reveal their knowledge gaps to them, for example, when they cannot complete a diagram box regarding a specific relation in the text (see Fig. 1, the last box on the right in the upper student diagram). In other words, the cue *box left empty* (i.e., omission error) partially predicts comprehension test performance (van de Pol et al. 2020). So, if students would base their monitoring judgement on this cue, their monitoring accuracy would improve.

van Loon et al. (2014) showed that completing pre-structured diagrams about the causal relations in texts was effective for improving students' monitoring accuracy of their text comprehension when done at a delay. That is, monitoring accuracy only improved significantly when

See your own and a correct diagram below.

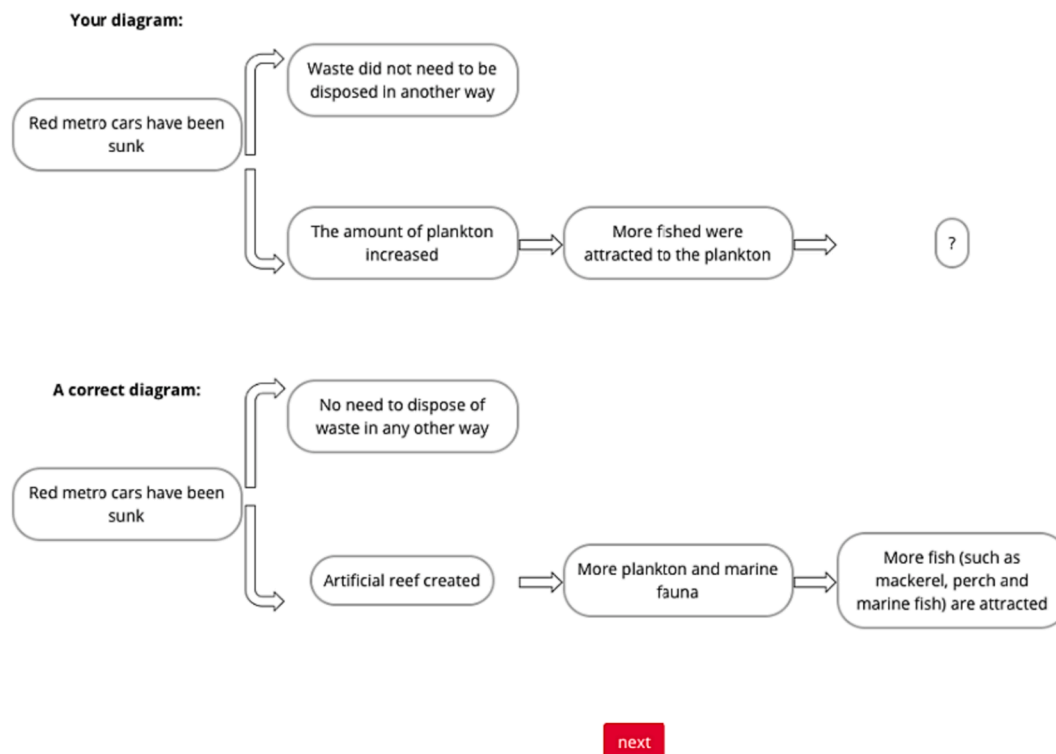


Fig. 1. Diagram Study Screen Example (Translated from Dutch).

first all texts were read and subsequently all diagrams completed, compared to when diagrams were completed immediately after reading a text. van de Pol et al. (2019) replicated the finding that monitoring accuracy was significantly higher after delayed diagramming than after a filler task. In both studies, the cues *number of diagram boxes completed correctly* and *number of diagram boxes left empty* appeared to be most diagnostic of students' test scores. These findings suggest that diagrams helped students to focus more on diagnostic cues, and that this, in turn, improved their monitoring accuracy.

However, even when using generative tasks, monitoring accuracy was found to be far from perfect and could be further improved (Prinz et al., 2020b; van de Pol et al., 2020). One reason for this might be that students tend to believe their answer is (at least partially) correct whenever something comes to mind regardless of its quality (Koriat, 1993), which can lead to overestimations of their comprehension, and hence, inaccurate monitoring judgements (Dunlosky et al., 2011; Rawson & Dunlosky, 2007; Waldeyer & Roelle, 2020; Zamary et al., 2016). In the case of the diagramming intervention, it could be that students were unable to distinguish between correctly and incorrectly completed diagram boxes (van de Pol et al., 2020). For example, van de Pol et al. (2020) reported that students with low and high monitoring accuracy did not differ in the number of boxes they completed, or in their scores on a subsequent comprehension test. However, students with low monitoring accuracy made more commission errors, whereas students with high monitoring accuracy completed more boxes correctly. These findings underline the negative relation between commission errors during diagramming on monitoring accuracy. Possibly, students use the number of completed boxes as a cue, but the cue is only diagnostic of test performance when those boxes are completed correctly (van de Pol et al., 2020). Therefore, one way to make the diagramming intervention more effective for improving monitoring accuracy could be to provide feedback in the form of a correctly completed standard to reduce the negative effect of relying on non-diagnostic information.

1.1.1. Diagramming and correct standards for improving monitoring accuracy

To the best of our knowledge, the effect of providing feedback after diagramming on monitoring accuracy has not yet been investigated. Nonetheless, previous research has reported positive effects of providing feedback-standards in combination with other generative activities for improving higher education students' (Dunlosky et al., 2011; Waldeyer & Roelle, 2020) or middle school students' (Lipko et al., 2009) monitoring judgements. The underlying assumption is that standards enable learners to assess the quality of their answers, which can lead to more accurate monitoring judgements (Rawson & Dunlosky, 2007). This assumption is also supported by studies that implemented recalling key definitions of learned concepts (Dunlosky et al., 2011) or delayed keyword generation (Waldeyer & Roelle, 2020) as generative tasks with standards as feedback. For instance, when students learned concept definitions or keywords and were tested on their recall of those definitions or keywords, providing feedback in the form of a correct answer - or a standard answer split into its idea units - improved students' self-assessment of their recall performance (Dunlosky et al., 2011) and relative monitoring accuracy (Waldeyer & Roelle, 2020). Hence, implementing standards could be a promising addition to generative task interventions that aim to improve monitoring accuracy.

However, only the study by Waldeyer and Roelle (2020) investigated the effect of providing standards for improving monitoring accuracy of text comprehension (other studies focussed on recall of key term definitions) during a generative task. Note, that previous research has shown that providing standards in the form of worked examples can improve monitoring accuracy and regulation of problem-solving tasks (e.g., Baars et al., 2014). Our standard can also be seen as a worked example of a correctly completed diagram. To the best of our knowledge, there is no research that employed standards of causal diagramming to improve monitoring accuracy of text comprehension.

Furthermore, all studies mentioned above implemented a delayed block design in which all key term definitions or keywords were studied first and only later all definitions or keywords were generated (in a block). Yet, at school or university, students often study only one text at a time. Consequently, they have to accurately assess their understanding of that particular text immediately after text reading in order to decide whether they have to restudy (parts of) the text or not. It therefore seems promising to further investigate the role standards could play in immediate generative task designs.

1.1.2. Immediate diagramming and standards for improving monitoring accuracy

The two studies that found an improvement of monitoring accuracy after diagramming only found this with a delayed design, when students first read all texts, then completed all diagrams and thereafter made all monitoring judgements (van de Pol et al., 2019; van Loon et al., 2014). Comparing delayed and immediate diagramming, van Loon et al. (2014) found that with an immediate design, there was a numerical, but not significant improvement compared to a no-diagramming control condition. The limited success of immediate designs is often explained by how situation models (Kintsch et al., 1990) of a text are formed (e.g., Anderson & Thiede, 2008; Prinz et al., 2020b). When performing generative activities at a delay, irrelevant factual (or surface-level) information will have naturally decayed, and students have to rely on information from long-term memory. Retrieving information from long-term memory more closely resembles the situation at the time of the test (Thiede et al., 2010). This means that the tasks (e.g., diagrams) will be completed with the same information that is available for answering the test question. However, when performing immediate generative activities after text reading, students can presumably rely on information still present in their working memory, which often also contains (irrelevant) factual knowledge that may lead to commission errors (i.e., faulty diagram boxes in case of a diagram task). Thus, immediate generative activities can lead students to overestimate their test performance (see Griffin et al., 2008; Thiede et al., 2010 or Prinz et al., 2020a,b for a more elaborate rationale).

However, arguably, when adding feedback in the form of a standard to the immediate design, the standard could function as a filter for irrelevant information. For example, if a student reads a text and immediately completes a causal diagram with irrelevant factual information from the text, the standard will provide the student with the opportunity to identify the irrelevant information, while also providing a restudy opportunity of the most relevant causal relations in the text. Thus, completing a diagram immediately after reading a text, might be effective for improving monitoring accuracy when combined with correct diagram-standards as feedback to help identify the most important causal relations of a text and to make students aware of potential mistakes. Therefore, in the present studies we asked for monitoring judgements (from now on *judgements of learning*; JOLs) immediately after reading each text, had students complete a diagram, provided them with a correct diagram-standard, and had them make another JOL immediately after completing each diagram. This allowed us to investigate if JOLs were adapted after diagramming and/or receiving a standard (and if so how).

The choice of an immediate diagramming design in combination with a standard has further implications for the measurement and conceptualization of monitoring accuracy in terms of relative and absolute measures. The studies by van Loon et al. (2014) and van de Pol et al. (2019), established relative accuracy of monitoring judgements *across texts*. Relative monitoring accuracy indicates whether students made a distinction in their judgements between texts that were well and less well understood (Griffin et al., 2019), as indicated by their comprehension test performance. This measure is therefore often used when monitoring judgements concerning different texts directly follow one another. However, it does not say anything about how accurate their judgement about their understanding of each particular text was (Griffin

et al., 2019), while students often need to make a restudy decision of one text directly after reading. In this case, calculating absolute accuracy seems more informative (Dunlosky & Rawson, 2012), as it represents the deviation between the monitoring judgement and comprehension test performance on the corresponding text. Nevertheless, for the sake of completeness and comparability, we report both absolute and relative measures, as recommended by Dunlosky and Thiede (2013).

1.2. Immediate diagramming and standards for improving text comprehension

Next to improving monitoring accuracy, simply providing students with a correct diagram might also improve their text comprehension. McCrudden et al. (2007) found that studying a correct causal diagram, without asking students to generate a diagram themselves, improved text comprehension. However, integrating a generative task before providing a standard as feedback seems crucial for monitoring accuracy, as Redford et al. (2012) found that only studying a completed concept map, without preceding generative task, did not improve monitoring accuracy. Furthermore, studies implementing generative tasks immediately after reading each text, such as immediate summarization (Anderson & Thiede, 2008, Thiede & Anderson, 2003), immediate diagramming (van Loon et al., 2014), or even concurrent concept mapping with reading a text (Thiede et al., 2010), also reported a direct improvement of text comprehension. Thus, our intervention with immediate diagramming followed by studying a correct diagram-standard might improve both monitoring accuracy and text comprehension (Hacker & Bol, 2019).

1.3. Gaining insights into students' cue use

Next to investigating whether feedback in the form of a standard would improve monitoring accuracy in an immediate diagramming design, it is important to establish *how* feedback improves monitoring accuracy. More specifically, investigating which cues students identify and use from comparing their diagram with the correct standard can help to render future interventions more adaptive. For example, instead of displaying a full feedback standard, the intervention could selectively guide students' attention to the diagnostic information to improve their monitoring accuracy and text comprehension.

Empirically measuring and examining which cues students base their JOLs on is challenging. There are two approaches in the literature, namely, an indirect approach of estimating cue utilization (e.g., Thiede et al., 2019; van Loon et al., 2014; van de Pol et al., 2019), and a direct approach of asking students directly which cues they used for their JOLs (e.g., Bol et al., 2010; Thiede et al., 2010). The indirect approach measures for cue use are operationalized as the relation between the quantitative value of a cue (e.g., number of correctly completed boxes) and a student's JOL. A limitation is that even when finding a significant relation between the presence of a cue and a student's JOL score, it is still unclear whether the student used this cue deliberately for making the JOL. Hence, for a researcher to understand why students made a particular JOL, a more direct approach of measuring cue utilization might be preferable.

This can be achieved, for instance, by asking students to think aloud while making JOLs, or by asking them directly what information or cues they used for making their JOL (Bol et al., 2010; Dinsmore & Parkinson, 2013; Händel & Dresel, 2018). However, concurrent reporting can be hard for students, especially under conditions of high cognitive load (van Gog, 2006), and retrospective reporting has the drawback that it can be hard for students to remember what they were thinking while performing the task. One way to overcome these obstacles is to use *cued retrospective reporting* (van Gog et al., 2005). The idea behind this technique is to record students' gaze via eye tracking while they complete a task, and then, after task completion, use replays of the recorded gaze patterns as a cue for retrospective verbal reporting. Students thus review

a video of their task performance, with their gaze location shown as dots or circles overlaid on the task, while they are asked to report what they were thinking during task performance. This technique has been employed successfully to investigate cognitive processes underlying visual tasks, for example, electrical circuits problem-solving (van Gog et al., 2005), evaluation of internet sources during web search (Brand-Gruwel et al., 2017), or multimedia learning (see van Gog & Scheiter, 2010, for an overview). van Gog et al. (2005) showed that participants who saw their recorded gaze patterns during the retrospective verbalization phase were able to report more executed actions and made more metacognitive statements compared to a group that did not receive the gaze display during retrospective verbalization. Cued retrospective reporting could, therefore, be a promising method to study how a correct diagram (as standard) is processed and what cues students potentially derive from their own diagrams and from the provided standard.

In addition to using the gaze patterns as input for verbal reports, the recorded eye movements can also be used to analyse students' processing of the diagram-standard. The gaze patterns could provide further insights into potential cues that students inferred from the comparison to their own diagram, which they may or may not have verbalized. For instance, gaze patterns could show that a student fixated commission errors longer than correct boxes, even if the student hardly mentioned commission errors explicitly during retrospective reporting. Also, how students compare the standard to their own diagram can give insight into cues students focused on. Holmqvist et al. (2011) indicate that longer fixation durations on certain areas can reflect the recognition of incorrect features (e.g., commission errors in our case), and that more transitions between areas can reflect awareness of the importance of an area (e.g., in our case, thoroughly comparing mismatching diagram boxes while spending less time on matching boxes; see Holmqvist et al., 2011 for a deeper discussion of the measures).

1.4. The present studies

The first aim of the present studies was to investigate the effect of diagramming, receiving a correct diagram (as feedback standard), and their combination on students' monitoring accuracy and text comprehension (Study 1). Secondly, we aimed to explore students' processing of the standard along with their own diagram and use of cues generated through diagramming by means of direct process measures such as eye-tracking and cued retrospective verbal reporting (Study 2).

2. Study 1

In this experiment, we compared students' absolute monitoring accuracy and text comprehension after (1) a no-diagramming filler task (No-Diagram; control), (2) receiving a correct diagram to study immediately after studying a text without completing a diagram (Standard-Only), (3) completing a causal diagram immediately after studying a text (Diagramming-Only), or (4) completing a diagram immediately after studying a text and then receiving a correct diagram as feedback standard (Diagramming + Standard). We included the condition that only studied a diagram-standard because prior research has shown that this can improve text comprehension (McCrudden et al., 2007). Therefore, the standard-only condition would enable us to disentangle the combined effect of diagramming and receiving a standard on both monitoring accuracy and text comprehension. To determine the influence of the experimental tasks (e.g., studying their own diagram next to a correct standard) on JOLs, participants were asked to make two JOLs: first immediately after studying the text, and second after the experimental task. We aimed to answer the following research questions:

RQ 1.1 Does (i) immediate diagramming (versus no diagramming), and (ii) receiving a (correct diagram-) standard (versus not receiving it) improve absolute monitoring accuracy, and (iii) does the effect of receiving a standard on monitoring accuracy depend on whether or not students first engaged in diagramming?

Based on van Loon et al. (2014), who found that delayed but not immediate diagramming improved relative monitoring accuracy, we did not necessarily expect a main effect of diagramming-only on absolute monitoring accuracy. However, we did expect that being able to compare one's own diagram to a standard would improve (absolute) monitoring accuracy compared to all other conditions.

RQ 1.2 Does (iv) immediate diagramming or (v) receiving a diagram-standard improve text comprehension?

Based on van Loon et al. (2014), we expected that immediate diagramming would improve text comprehension. Furthermore, based on research on studying completed causal diagrams (McCrudden et al., 2007), we expected that displaying a diagram-standard (also without previous self-diagramming) would have a positive effect on text comprehension.

RQ 1.3 Do students change their JOLs from before to after the experimental tasks in the different conditions (i.e., Diagramming + Standard, Diagramming-Only, Standard-Only, or the picture matching filler task in the No-Diagram condition) and does the degree of change differ between conditions?

This will provide insight into the potential effect of the experimental conditions on the height of students monitoring judgments, independent of their accuracy (which also depends on the test scores).

2.1. Methods

2.1.1. Participants and design

Based on an a-priori power analysis,¹ 80 participants ($M_{age} = 20.00$; $SD_{age} = 1.69$; 35% female) from the Netherlands and Belgium were recruited via the Prolific online platform (<http://www.prolific.com>). To get a sufficiently large sample older than 18 (minimum participant age allowed in Prolific), that would not be too familiar with reading complex (academic) texts, we aimed to sample people without a completed bachelor's degree. Hence, Prolific settings were adjusted to only allow Dutch native speakers under the age of 23,² with a secondary education degree (but not higher), to participate.³ Fifteen other (additional) participants started the experiment but stopped at the instruction part and were thus not included in the sample (this occurred in all conditions: No Diagram: $n = 5$; Standard Only: $n = 2$; Diagramming-Only: $n = 4$; Diagramming + Standard: $n = 4$). Data from one participant in the Standard-Only condition was excluded due to an unrealistically short completion time of under two minutes. So, the final sample consisted of 79 participants ($M_{age} = 20.00$, $SD = 1.68$, 35.44% female). Ethical approval for this study was obtained from the first author's institute and participants gave informed consent before starting the experiment. Participants who completed the experiment were compensated with 10 €.

The study had a 2x2 between-subjects design with *diagramming* (yes/no) and receiving a *standard* (yes/no) as factors. Participants were randomly assigned to one of four conditions in which they first read a text and then: performed a filler-task (*No-Diagram control condition*; $n = 19$), studied a (correct diagram-) standard only (*Standard-Only condition*; $n = 21$), completed a diagram and studied it (*Diagramming-Only*; $n = 20$), or completed a diagram and received a standard to study (*Diagramming*

¹ The power analysis was conducted with the SPA-ML software (Moerbeek & Teerenstra, 2016) for multilevel analyses, using effect size $d=0.40$ and ICC = 13.4% (on student level), based on Van De Pol et al. (2019).

² Students in the Netherlands finish a bachelor's degree around the age of 22.

³ Note that we first ran the experiment with thirteen participants to ensure the functionality of the software; at this point the age limit setting had not yet been correctly implemented. Because everything worked well, the experiment was filled up until 80 participants, with the age limit setting. In the first 13 participants, three students were older than 23. We ran the analyses with and without the older participants, and since this resulted in the same outcome patterns, we did not remove them from the sample.

+ Standard; $n = 19$). See Fig. 2 for an overview of the conditions.

2.1.2. Materials

The experiment was implemented online on the Gorilla Experiment Builder (<http://www.gorilla.sc>). We used the same texts, diagrams, and pictures as van Loon et al. (2014) and van de Pol et al. (2019).

Instruction Videos and Practice Trials. Four instruction videos (one for each condition) were created to give participants an example of the experimental procedure for one trial in each condition. The videos contained the following sequence: First, an example text was displayed; then they observed the first JOL being filled out; then, depending on condition, they observed a diagram completion and/or diagram study task or a picture matching filler task being completed; the second JOL that was filled out; and the video ended with completion of the comprehension test (see Fig. 2 for the trial design per condition). The example text in the instruction videos was a short summary of the first scene of *Alice in Wonderland* (Carroll, 2015). Additional instructions (e.g., "Each trial begins with the task to read a text" appearing above the text) were displayed in a blue, bold font next to the experimental instructions (in black font) on each screen. The test format (i.e., requiring four causal relations as response to one question) was additionally explained next to the first JOL example. The instruction videos could not be paused and lasted between 102 s (Standard-Only) and 173 s (Diagramming + Standard). See the instruction video of the Diagramming + Standard Condition in the supplementary materials and Appendix A for a translation of the additional instructions during the video.

The video was followed by a practice trial depicting trials in the respective condition and was comparable to the six actual study texts in terms of text length, diagram structure or picture matching filler task, and test question.

Texts. The study materials consisted of six expository texts in Dutch (see Appendix B for two translated examples; van Loon et al., 2014). The texts contained 158–178 words which were displayed in a single paragraph. They included words or expressions signalling four causal relations in the texts, through the Dutch equivalents of *therefore*, *for this reason*, or *this meant that*. The topics of the texts were: "Botox", "Sinking of metro cars", "Concrete constructions", "The Suez Canal", "Money does not bring happiness", and "Music makes smart".

Judgements of Learning (JOLs). Students were asked to predict their test scores by clicking one of five buttons (response options ranging from 0 to 4) as an answer to the question: *How many points do you think you would score on a test question about the text [text name]?* They appeared after each text (JOL₁) and after the experimental task (JOL₂).

Diagrams and Filler Task. The diagrams were composed of five boxes (see Fig. 1 for an example). In the diagrams that students had to complete (i.e., in conditions Diagramming-Only and Diagramming + Standard), the first box was filled out (i.e., given box) and the other four boxes were empty for students to complete. In the correct diagrams that students had to study (i.e., conditions Diagramming + Standard and Standard-Only), all boxes were already completed. The instruction for the diagram completion conditions was *Complete the diagram for the text [text name]. If you cannot complete a box, insert a "?"*. In the Diagramming-Only condition, the instruction for the diagram study phase was *Below, look at your own diagram once more*. In the Diagramming + Standard condition, the instruction was *Below, examine your own and a correct diagram* (see Appendix C for a translated trial example).

Participants in the control condition did not receive any diagrams but instead completed a picture-matching filler task, in which two pictures related to the text/trial topic (e.g., picture of the Suez Canal after text about the Suez Canal) were displayed next to one another with the instruction to count and state all differences between the pictures within 90 s (a countdown was displayed). The instructions were: *On the next screen, you will see two pictures that are slightly different in a few places. How many differences are there? You have 90 s to find the differences and give your answer.*

Text Comprehension Test. Text comprehension was measured with

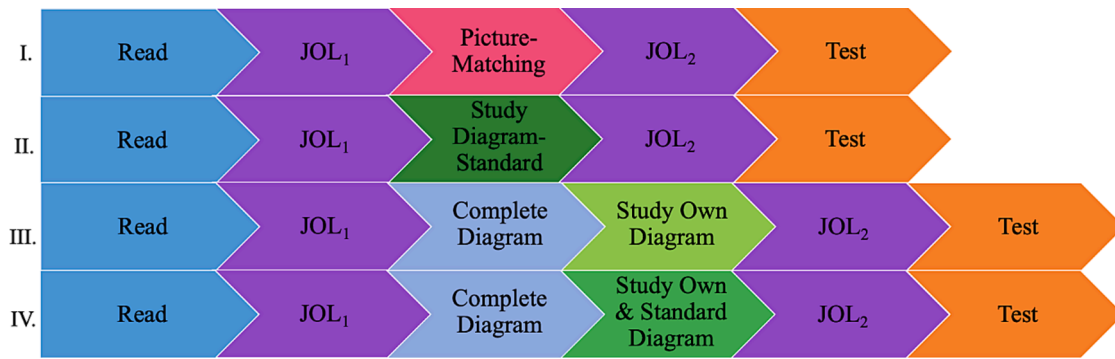


Fig. 2. Trial Design Across Conditions. *Note.* JOL: Judgement of Learning; Conditions: I. No-Diagram (control), II. Standard-Only, III. Diagramming-Only, IV: Diagramming + Standard.

a test about each text consisting of one question asking to describe the four causal relations from each text. For example, in case of the Botox text (Appendix B) the question was: “Botox blocks the tightening signal between the nerves and muscles in our face. What are the consequences of this?” All text specific questions were followed by the instruction: “Try to give as complete an answer as possible, mentioning four relations in it. If you don’t know anything you can fill in a question mark. Good luck!” Possible scores ranged from 0 to 4 as students received one point for each correctly mentioned causal relation. Responses were scored as being correct if they were part of a correct logical reasoning to answer the test question. That is, simply naming any propositional representations (reflecting a processing level preceding the information integration on the situation model level according to Kintsch et al., 1990) named in the text, was not enough to answer our test questions.

2.1.3. Procedure

Participants started the experiment in Prolific (the online recruiting platform) and were redirected to the online experiment in Gorilla via a link. The experiment was only accessible from a laptop or computer for comparability across participants. Once in Gorilla, participants received a detailed information letter about the study and were required to provide consent for participation in order to proceed. Subsequently, participants completed a checklist by confirming to be seated at a quiet place with a proper internet connection, where they would not be disturbed for the next 40 min. Furthermore, participants were asked to not take notes during the experiment. Then, the instruction video played, explaining the procedure for participants’ assigned condition. After the video-instruction, participants completed a practice trial and

subsequently proceeded to the six experimental trials, where they read a text, made a first judgement of learning (JOL₁), completed a task based on condition, made a second judgement of leaning (JOL₂), and ultimately completed a comprehension test (see Fig. 2). All trials were self-paced and the order was randomized across participants. Including only the time spent on the instruction and tasks of each condition, participants spent on average 24.83 min (SD = 8.39) in the No-Diagram control condition, 17.84 min (SD = 6.32) in the Standard-Only, 34.03 min (SD = 11.94) in the Diagramming-Only, and 29.81 min (SD = 12.27) in the Diagramming + Standard condition.

2.2. Data analysis

The first author and an assistant coded 20% of the diagrams and comprehension tests and reached an inter-rater reliability (weighted squared kappa; κ) of 0.94 for the diagrams and 0.88 for the tests. The first author then coded the remaining data. Note that if the causal relations of content were true, boxes were coded as ‘correct’, even if they were not at the matching positions in the diagram-standard. Imagine, for example, a standard with Response A → Response B → Response C → Response D, if a student’s diagram contained Response A → C → D, three out of four boxes would have been scored correctly, regardless of the potentially mismatching positions of Response C and D if no box was left empty for the missing Response B.

Absolute monitoring accuracy (i.e., deviation) was calculated as the unsigned difference between a participant’s JOL₂ and the respective comprehension test score per trial. The range was thus 0–4, with 0 meaning fully accurate (i.e., no deviation at all) and 4 meaning fully

Table 1
Descriptive Statistics of Monitoring Accuracy, Test Scores, and JOLs.

	Monitoring Accuracy		Test Scores (0–4)	JOL ₁ (0–4)	JOL ₂ (0–4)
	Absolute (0–4)	Relative ¹ (-1–1)			
	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)
Study 1					
Conditions					
Diagramming + Standard	0.68 (0.61)	0.38 (0.68)	3.65 (0.56)	2.27 (0.91)	3.18 (0.79)
Diagramming-Only	0.98 (0.81)	0.25 (0.56)	3.08 (0.87)	2.52 (0.86)	2.58 (1.05)
Standard-Only	0.75 (0.67)	0.40 (0.68)	3.23 (0.89)	2.34 (0.90)	2.94 (0.86)
No-Diagram (Control)	1.02 (0.91)	0.19 (0.61)	2.90 (1.02)	2.57 (0.87)	2.40 (0.96)
Predictor-Levels (for analyses)					
Diagramming	0.84 (0.73)	0.30 (0.61)	3.35 (0.79)	2.40 (0.89)	2.87 (0.98)
No Diagramming	0.88 (0.80)	0.29 (0.65)	3.07 (0.97)	2.45 (0.89)	2.68 (0.94)
Standard	0.72 (0.64)	0.39 (0.67)	3.43 (0.78)	2.31 (0.90)	3.05 (0.83)
No Standard	1.00 (0.86)	0.22 (0.58)	2.99 (0.95)	2.54 (0.86)	2.49 (1.01)
Total	0.88 (0.79)	0.30 (0.62)	3.19 (0.93)	2.42 (0.89)	2.77 (0.96)
Study 2					
Diagramming + Standard	0.79 (0.79)	0.44 (0.69)	3.64 (0.65)	2.21 (0.80)	3.02 (0.85)

Note. None of the outcomes significantly differed comparing the Diagramming + Standard conditions of Study 1 and 2.

¹ Gamma correlations of ten participants in Study 1 and eight participants in Study 2 could not be computed because of ties.

Table 2
All Resulting Multilevel Model Parameters.

Models	<i>B</i>	<i>S.E.</i>	<i>CI</i>	<i>p</i>	<i>b</i>	
Hypothesized Model RQ1						
Monitoring Accuracy ~						
Diagramming (D)	-0.04	0.12	-0.27 – 0.20	.764	0.05	
Receive Standard (RS)	-0.27	0.12	-0.50 – 0.04	.023*	0.35	
Interaction (D*RS)	-0.03	0.17	-0.36 – 0.29	.846	0.04	
Simplified Model/Final Model RQ1						
Monitoring Accuracy ~						
Diagramming (D)	-0.05	0.08	-0.21 – 0.11	.531	0.07	
Receive Standard (RS)	-0.28	0.08	-0.44 – -0.12	.001*	0.37	
Relative Accuracy – RQ1						
Monitoring Accuracy (gamma correlations) ~						
Diagramming (D)	0.06	0.21	-0.35 – 0.47	.780	0.09	
Receive Standard (RS)	0.21	0.22	-0.22 – 0.65	.334	0.34	
Interaction D*RS	-0.05	0.29	-0.62 – 0.52	.852	-0.08	
Hypothesized Model RQ2						
Text Comprehension ~						
Diagramming (D)	0.18	0.15	-0.12 – 0.47	.237	0.20	
Receive Standard (RS)	0.32	0.15	0.03 – 0.61	.029*	0.36	
Interaction D*RS	0.25	0.21	-0.16 – 0.66	.234	0.28	
Simplified Model/Final Model RQ2						
Text Comprehension ~						
Diagramming (D)	0.30	0.11	0.10 – 0.51	.004*	0.34	
Receive Standard (RS)	0.45	0.11	0.24 – 0.66	.001*	0.50	
Additional Exploratory Analyses: Pairwise Comparisons of Conditions						
Monitoring Accuracy ~	<i>B</i>	<i>S.E.</i>	<i>df</i>	<i>t</i>	<i>p</i>	<i>b</i>
DO – DS	0.30	0.12	74.37	2.54	.061	0.41
DS – SO	-0.07	0.12	75.08	-0.58	.938	-0.09
ND – DS	0.33	0.12	74.78	2.80	.032*	0.46
DO – SO	0.23	0.11	75.10	2.01	.192	0.32
ND – DO	0.04	0.12	74.79	0.30	.990	0.05
ND – SO	0.27	0.12	75.51	2.29	.110	0.37
Text Comprehension ~						
DO – DS	-0.57	0.15	74.52	-3.85	.001*	-0.73
DS – SO	0.43	0.15	75.07	2.89	.025*	0.54
ND – DS	-0.75	0.15	74.84	-4.96	<.001*	-0.95
DO – SO	-0.15	0.15	75.08	-1.01	.746	-0.19
ND – DO	-0.18	0.15	74.85	-1.18	.639	-0.22
ND – SO	-0.32	0.15	75.40	-2.19	.136	-0.41

Note. *B*: non-standardized effects, *S.E.*: standard error, *b*: standardized effects. * = significant ($p < .05$); DS: Diagramming + Standard, DO: Diagramming-Only, SO: Standard-Only, ND: No-Diagram (control). The reference level of the pooled predictors was No-Diagramming and Not-Receiving-Standard, and for the conditions No-Diagram. The analysis included 471 observations of 79 participants. Ten observations for relative monitoring accuracy were deleted during gamma correlations computation. The final model for RQ1 explained 11.0% of the observed variance, and the final model for RQ2 explained 23.1%. Alpha of pairwise comparisons were Tukey-corrected for multiple testing.

Table 3
Main and Interaction Effects for JOL-Instance and Condition on JOL-scores.

Predictors	B	S.E.	CI	p	b
JOL ₂	-0.17	0.10	-0.36 – 0.03	.089	-0.18
Diagramming-Only	-0.05	0.19	-0.43 – 0.33	.783	-0.06
Diagramming + Standard	-0.30	0.20	-0.68 – 0.09	.130	-0.32
Standard-Only	-0.23	0.19	-0.61 – 0.15	.233	-0.24
JoL ₂ × Diagramming-Only	0.22	0.14	-0.04 – 0.49	.101	0.24
JoL ₂ × Diagramming + Standard	1.07	0.14	0.80 – 1.34	<.001	1.14
JoL ₂ × Standard-Only	0.76	0.14	0.50 – 1.03	<.001	0.81

Note. The reference level of JOL-Instance was JOL₁ and of condition No-Diagram.

inaccurate (i.e., maximum possible deviation). For example, if a student made a JOL of 2 and subsequently scored 4 points on the test, the absolute monitoring accuracy would be 2 (i.e., $|2 - 4| = |-2| = 2$ as unsigned absolute difference). For completeness and comparability with other studies (e.g., van Loon et al., 2014; van de Pol et al., 2019), we also report relative monitoring accuracy, which is an indicator of monitoring accuracy across texts. Following the literature, it was operationalized as the Goodman and Kruskal's (1979) gamma correlation (range -1 to 1) between the students' JOLs and test scores for the six trials of the experiment.

Multilevel models were fitted to account for variance dependencies of multiple observations within each participant (six trials nested in one participant). The four conditions were pooled based on whether they included diagram completion or not (factor *diagramming*) and whether they included receiving a standard or not (factor *standard*). The first model contained monitoring accuracy as dependent variable and both factors diagramming and standard as predictors. It was checked whether the model containing the interaction predicted significantly more variance compared to the model containing only the main effects (Dalpiaz, 2022). The second model contained text comprehension as dependent variable and the same two main effects, and their interaction as predictor. The third model contained JOL as dependent variable and the predictors JOL-Instance (First/Second) and condition. Analogue to the first model, the interaction of JOL-Instance and condition was added in a second step. The effect sizes are provided as standardized estimates (*b*). Post-hoc pairwise comparisons were conducted with Tukey-tests to correct *p*-values for multiple testing. All analyses were conducted with R (R Core Team, 2022) in RStudio (Posit Team, 2022). All analysis scripts for Study 1 and 2 are accessible online (see Appendix D).

2.3. Results

The descriptive statistics related to the research questions are provided in Table 1 and all outcome parameters of the multilevel analysis are listed in Table 2. Of the JOLs, 82.28% diverged ≤ 1 point from the corresponding test scores, indicating that monitoring accuracy was very high in this sample.

2.3.1. Separate and joint effects of diagramming and receiving a standard (RQ1-RQ2)

As for monitoring accuracy (RQ 1.1), we found a main effect of receiving a Standard (i.e., the factor with the pooled conditions in which standards were provided; factor level: Yes; or factor level: No). There was no significant main effect of the factor Diagramming, nor –in contrast to our hypothesis– a significant interaction effect (Table 2). Adding the interaction between Standard and Diagramming to the simpler model without the interaction effect, did not predict significantly more variance than a model without the hypothesized interaction, $\chi^2(1) = 0.04, p = .843$, so the simplified model was interpreted. Even though the interaction effect was not significant, the pattern in the means displayed in Table 1 strongly suggests that the significant main effect of receiving a Standard (i.e., the factor with the pooled conditions)

was primarily driven by the Diagramming + Standard condition, which showed the highest monitoring accuracy (i.e., lower score = better). To investigate this possibility, exploratory follow-up analyses were conducted by pairwise comparing monitoring accuracy in the four conditions. The pairwise comparisons showed a significant difference between the No-Diagram control condition and the Diagramming + Standard condition, $t(74.78) = 2.80, p = .032, b = 0.46$, while none of the other comparisons were significant (see Appendix E for the outcomes of all pairwise comparisons).

Regarding text comprehension (RQ 1.2), there were—as expected—significant main effects of the factors diagramming and receiving a standard on text comprehension. Adding the interaction of the factors (i.e., the pooled conditions) to the model did not explain significantly more variance, $\chi^2(1) = 1.48, p = .224$, so the simplified model was interpreted. Participants who completed a diagram performed significantly better on the comprehension test than participants who did not, and participants who received a standard performed significantly better than those who did not.

Again, to further explore the mean patterns where the Diagramming + Standard condition displayed the highest text comprehension performance, additional exploratory analyses were conducted to disentangle the effects of the four single conditions. The pairwise comparisons showed that students in the Diagramming + Standard condition scored on average significantly higher on the text comprehension test than students in the other conditions (i.e., No-Diagram, $t(74.84) = -4.96, p < .001, b = -0.95$, Standard-Only, $t(75.07) = 2.89, p = .025, b = 0.54$, or Diagramming-Only, $t(74.52) = -3.85, p < .001, b = -0.73$), while there were no other significant differences between conditions (see Appendix E for all outcomes of the pairwise comparisons). This means that test performance did not simply increase based on access to more information as provided through the standard alone.

2.3.2. Do learners change judgements of learning after diagramming and/or a Standard? (RQ3)

The multilevel analysis with JOL scores as outcome of interest showed no significant main effect of condition, but there was a significant main effect of JOL-Instance (i.e., JOL₁/JOL₂), and a significant interaction of JOL-Instance and condition (see Table 3 for all main and interaction effects). The model with the interaction explained significantly more variance compared to the model without the interaction, $\chi^2(3) = 73.08, p < .001$. Post-hoc comparisons only showed significant differences between JOL₁ and JOL₂ ratings in the conditions Standard-Only, $t(865) = -6.38, p < .001$, and Diagramming + Standard, $t(865) = -9.22, p < .001$. Thus, the interaction effect indicates that receiving a standard increased JOL₂ ratings compared to JOL₁ ratings, whereas diagramming or the filler task did not change students JOL.

As one would expect (since only the text has been read at that point), pairwise comparisons showed that there were no significant differences between JOL₁ scores across conditions. Furthermore, JOL₂ scores of the Diagramming + Standard condition were significantly higher than JOL₂ scores of the No-Diagram control condition, $t(97.61) = -3.92, p = .004$ and almost significantly higher than JOL₂ scores of the Diagramming-Only condition $t(97.61) = -3.09, p = .051$. There were no other significant differences between JOL₁ scores across conditions. See Appendix F for all pairwise comparisons and Fig. 3 for a visualization of the JOL-mean scores across conditions.

2.4. Discussion Study 1

The aim of the first study was to investigate the effect of diagramming, receiving a standard, or a combination of both, on students' monitoring accuracy and text comprehension. In contrast to our expectation, there was no significant interaction effect between diagramming and receiving a standard. However, we did find a main effect of receiving a standard, and exploratory follow-up analyses

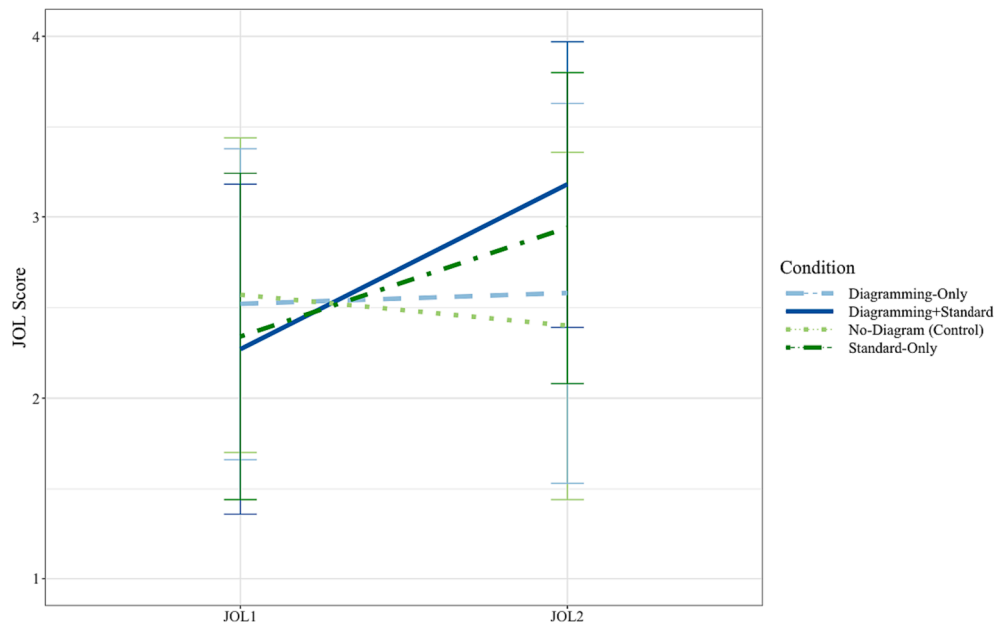


Fig. 3. Average Change JOL after Diagramming/Standard per Condition.

showed this was mainly driven by the combined Diagramming + Standard condition (which was the only condition to show significantly better monitoring accuracy compared to the No-Diagram control condition).

Not only monitoring accuracy, but also text comprehension significantly improved after a combination of diagramming and receiving a standard (but not by either diagramming or receiving a standard alone).⁴ These findings suggest that receiving a standard, especially after diagramming, has beneficial effects for monitoring and comprehension, and the data regarding the change in JOL ratings suggest that participants were aware of this. That is, they made significantly higher JOLs after studying the standard than after reading the text, while participants in the Diagramming-Only and in the No-Diagram control condition did not significantly change their JOLs.

3. Study 2

In Study 2, we employed eye tracking (Holmqvist et al., 2011) and cued retrospective verbal reporting (van Gog et al., 2005) to acquire insight into students' processing of the standards in relation to their own diagrams (RQ 2.1), and to get a better understanding of the cues they might gain from doing so (RQ 2.2). Thus, in Study 2, only the Diagramming + Standard condition of Study 1 was implemented. Matching the gaze data and diagram performance data (i.e., the diagram cues *correctly completed boxes*, *commission errors*, or *omission errors*) of each student, enables the analysis of how different fixation durations of, and transitions between, diagram boxes relate to the diagram cues.

3.1. Research questions

RQ 2.1 How do students process the standard in relation to their own diagram?

This was first explored by analysing eye-tracking data (i.e., total fixation duration on boxes, and transitions between the own diagram and the standard) for potential cues that could be extracted from the

⁴ Note that performance in this condition was very high and contrasts should be interpreted with caution.

diagram (i.e., correct boxes, commission errors, and omissions). These exploratory analyses can provide indications as to whether students were (presumably) aware of a certain cue (e.g., having committed a mistake) by focusing longer (i.e., longer total fixation duration) on the boxes they did not complete correctly than on boxes they did complete correctly, both in their own diagram and in the standard. In addition, we investigated how extensively students compared their own diagram to the standard (and boxes within one diagram) by counting fixation transitions within the own diagram and the standard, and between matching content or positions of the diagram boxes in students' own diagrams and the standard. Finally, in order to investigate how students processed the standard in relation to their own diagram, we analysed the protocols of the cued retrospective verbal reports by summarizing students' comments on 1) whether they compared the standard with their own diagram, 2) whether they identified matching and mismatching attributes of the two diagrams, and 3) whether they used the standard as preparation for the test that followed in the end of each trial.

Generally, in line with the findings of McCrudden et al. (2007), we expected that if students would focus on improving their text comprehension (as indicated by Study 1), they would study the standard, but not necessarily compare it with their own diagram. In other words, the own diagram might be perceived as less relevant for studying the most relevant inferences of the text depicted in the standard. This would then result in longer fixations on the standard compared to their own diagram and only few transitions. Yet, if they are focused on improving their monitoring accuracy, they would compare their own diagram with the standard (resulting in transitions between both diagrams).

Based on the findings of Study 1, where the combination of diagramming + studying a standard improved monitoring accuracy and text comprehension, students might also both compare their own diagram to the standard, and directly study parts of the standard that were not completed well in the own diagram and hence relevant to study for better text comprehension. In this case students would compare their own diagram with the standard (resulting in more transitions between the diagrams) but also study the standard longer compared to their own diagram (resulting in longer fixation durations on the standards).

RQ 2.2 Which cues do students infer from the two diagram types and use when making their monitoring judgements?

This was exploratorily investigated through the verbal protocols of

Table 4
Coding Categories for Think-Aloud Protocols (/Segments).

	Category	Description	Example Segments
Cue Type	Correct Box	Reference to a specific, correct box in one's own diagram	I had filled that in correctly. I thought that one was fine.
	Correct Relation	Reference to a specific correct relation in one's own diagram	I understood this connection. This sequence was correct.
	Incorrect Box Commission	Reference to a specific, incorrectly completed box (commission) in one's own diagram	I made a mistake here. That was not mentioned at all.
	Incorrect Box Omission	Reference to a specific box left empty (omission) in one's own diagram	I had not filled this in. I had put a question mark here.
General Content-Related Comparisons	Incorrect Box, Not-Specified	Reference to a mistake where the error type (commission or omission) was not specified	The second I did not know. I did not get it right.
	Incorrect Relation	Reference to a specific, incorrect relation in one's own diagram	The connection between... was not good. I saw that the two were reversed. ¹
	Mismatch-Identification	Reference to differing aspects between the diagrams, often involving signal words as <i>more</i> , <i>differently</i> , etc.	The correct diagram was more concise. I worded it differently.
	Match-Identification	General, not-cue-type-specific identification of overlapping attributes between the diagrams	My diagram was already quite similar to the correct diagram. The two diagrams were quite comparable.
	Own + Standard	Comparing own and standard diagrams	Here I am comparing very much. Then I went back and compared.
Processing-Related Statements	Own	Statements/strategies related to the inspection of one's own diagram	I had first looked at what I thought.
	Standard	Statements/strategies related to the inspection of the standard diagram	I was always going to read the correct diagram first. So that is why I immediately looked at the correct diagram.
	Other/General processing statements	Processing statements that were not clearly attributable to one of the two diagrams	I did not really look at that. I mainly focused on the last three. I always went from left to right. Here I am doing a final check.
Standard Diagram	Test Preparation	Using the standard diagram as preparation for test/ learning from standard diagram	I really tried to remember this. I saw that I could shorten my answers a bit. I was going to repeat the connections one by one before the test.
	(Meta-)Comments	(Meta-)Commenting on standard diagram answers/content	I didn't understand what they meant by 'cramped'. This did not make sense to me. I was surprised that it was different.
Own Diagram	(Meta-)Comments	(Meta-)Commenting on own diagram answers/content	Because I had written down 'paralysed'. Whereas, in the end, I had just said too much.
Context-Related	Previous Trials	Comparison to/with previous trials	I had prior knowledge of this subject.
	Topic/Text-Related	General statements regarding the topic or text of the trial	I had read the text carefully. I had difficulty in remembering the text.
Judgements of Learning	Metacognitive judgements	Statements regarding metacognitive judgements	I was confident. I was not very sure while filling in my own diagram.
	Judgements of Learning	Choosing scores	That is why I had chosen a 3. I have indeed pressed four.
Irrelevant for Analysis	Emotions	Statements related to feelings or emotions	I was proud. I was disappointed.
	Other	Statements that were not attributable to any of the previous categories	I have to think. What was I thinking here? Oh, that was quite a trip.

¹ Note that a reversed order in a causal diagram is incorrect by definition.

Table 5
Descriptive Statistics of Total Fixation Durations per Diagram and Cue Type.

Diagram Type	No. of (fixated) diagram boxes in eye-tracking data/ No. of diagram boxes in performance data	Average Total Fixation Duration of Single Box Per Cue Type (seconds)	% Total Fixations on Screen relative to screen duration
Cue Type		<i>M (SD)</i>	<i>M (SD)</i>
Own Diagram			21.59 (10.82)
Correct	322/352	1.29 (1.11)	16.85 (11.50)
Commission	73/80	1.21 (0.79)	2.99 (4.49)
Omission	21/31	0.40 (0.33)	0.45 (1.61)
(Given Box)	65/117	0.57 (0.64)	1.31 (2.05)
Standard			61.42 (14.58)
Correct	351/352	2.93 (2.32)	40.47 (17.24)
Commission	79/80	4.53 (3.93)	11.44 (17.31)
Omission	31/31	3.09 (2.78)	3.76 (8.44)
(Given Box)	115/117	1.42 (1.63)	5.75 (4.61)
Outside Diagrams/AOIs		4.51 (6.30)	16.99 (10.70)

the cued retrospective reporting by analysing how often participants named diagram cue types (e.g., “*This box was incorrect*” as an indication of a commission error) during the diagram study screen, and which information they mentioned as basis for their monitoring judgement on the JOL₂ screen (e.g., *I had two boxes incorrect, [so I chose a score of 2]*).

3.2. Methods

3.2.1. Participants and design

Twenty-two participants ($M_{age} = 20.12$; $SD_{age} = 1.41$; 77.27% female) were recruited online and via flyers at the first author's university with the same inclusion criteria as in Study 1. Only participants with normal or corrected-to-normal vision could participate. All participants completed the Diagramming + Standard condition of Study 1. Gaze data of two participants had to be excluded as their validation values exceeded 1.3 degrees of the visual angle, the maximal measurement inaccuracy that could be accounted for by the experimental set-up. The mean accuracy of the included gaze data was acceptable ($M = 0.55^\circ$, $SD = 0.18$). The tracking ratio of the included gaze data was over 75%. Performance data of all 22 participants was included. However, for three participants, only five instead of the six trials could be analysed because of technical issues (i.e., experimental software crash). All participants gave their informed consent prior to the study and were compensated based on their own preference with either 10 € or study credits.

3.2.2. Materials and measures

Introduction Video, Texts, Diagrams, JOLs, Test. The same materials (i.e., instruction video, texts, diagrams, tests, and overall trial-set-up) as in the Diagramming + Standard condition of Study 1 were

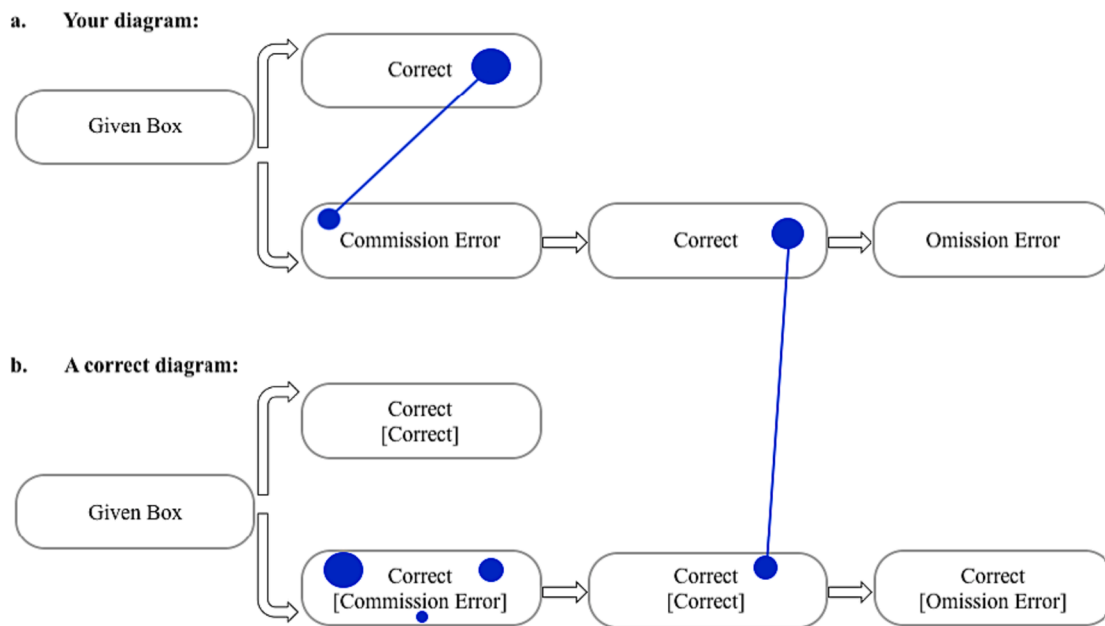


Fig. 4. Visualization of Gaze Measures per Cue Type per Diagram Type.

Note. Cue types (*Correct*, *Commission Error*, *Omission Error*) are depicted for both diagram types: *Own Diagram* (upper panel a.) and *Standard* (lower panel b.). All *Correct* boxes in the standard correspond to a certain cue type in the own diagram, depicted in square brackets. The blue dots in the first lower box in panel b. reflect fixation durations of different lengths (the larger the size the longer the duration) on a (correct) box in the standard that corresponds to the position of the own diagram box containing a commission error. The connected blue dots depict fixation transitions (between a commission error and a correctly completed box) *within* the own diagram (arrow in panel a.), and *between* boxes of the own diagram and standard (line between panels), see Appendix H for a screenshot of a real transition depiction. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

employed (also implemented on the Gorilla Experiment Builder; <https://www.gorilla.sc>).

Performance, Monitoring, and Indirect Cue Measures. JOLs, monitoring accuracy, and text comprehension were operationalized as in Study 1. Cue values were defined as the number of occurrences of a cue type within one diagram (e.g., cue value of correct boxes would be three, if three boxes in one diagram were completed correctly). The measures of cue diagnosticity and cue utilization were computed as Pearson correlation coefficients across all trials per participant between cue values and test scores (cue diagnosticity), and JOLs and test scores (cue utilization), respectively.

Apparatus and Set-Up. Eye movements were recorded at a sampling rate of 250 Hz with an SMI RED250 eye tracker (SensoMotoric Instruments GmbH, 2017) in binocular mode, and the study was implemented in the eye-tracker-associated software SMI Experiment Center (Version 3.7; SensoMotoric Instruments GmbH, 2017), from which the link to the online study in [Gorilla.sc](https://www.gorilla.sc) was opened. A chin rest stabilized the participant's head movements and facilitated gaze directions to the centre of the stimulus screen (22-inch monitor; 1680 × 1050 pixels) at a viewing distance of 70 cm.

AOI Definitions. Areas of Interest (AOIs; sometimes also referred to as Regions of Interest) were defined around all single diagram boxes, in such a way that, even with adaptive actual diagram box heights due to entered number of words, there would be a minimal distance of 60px around all single diagram boxes when about 25 words would be entered in one diagram box. AOI dimensions ranged between 205x161 and 242x161 pixels (see Appendix G for an example) across diagrams to fit the different diagram formats. Within trials, AOIs had the same dimensions for the own and standard boxes with no overlapping borders.

Total Fixation Duration on AOIs and Number of Transitions. Using the SMI high-velocity event detection algorithm with a threshold of 40°/s for detecting saccades, fixations were defined at a minimum of 50 ms (which is a normal cut-off; Holmqvist et al., 2011). The number of

fixated AOIs with a particular feature (i.e., diagram boxes with commission errors) is reported in Table 5, along with the actual number of occurrences (e.g., the total number of diagram boxes with commission errors) from which it can be subtracted how many diagram boxes were displayed on the screen but not fixated. Note that this is summed across participants (i.e., the maximum number of boxes that could be fixated across all participants and trials was 23,400 boxes over 60 different AOIs: 5 boxes in the own diagram + 5 boxes in the completed diagram = 10 boxes per participant per trial* 20 participants = 200 boxes* 117 trials (i.e., minus drop-out) = 23,400).

Fixation durations were summed up per AOI and reported as average total fixation duration (in s) on an AOI for each cue type (i.e., correct, commission, omission; and for completeness: given boxes) in diagram type (own and standard). That is, the total fixation duration on a single diagram box corresponding to one cue type (e.g., a commission error) in one diagram type (e.g., the own diagram) was averaged across the total fixation duration of all other single diagram boxes corresponding to the same cue in the same diagram (e.g., all other boxes with commission errors in all other own diagrams; see Fig. 4). Because the average number of words per cue and diagram type is likely to differ between own diagrams and the standard, this is also reported below to better allow for interpreting possible fixation duration differences.

Additionally, percentages of total fixation duration for each cue type and diagram type were computed relative to all fixation durations during screen presentation, including fixations that were not directed to one of the diagrams (e.g., dividing the sum of all fixation durations of all correct boxes in the own diagram by all other fixation durations on the screen). Note that the measure of average total fixation durations in seconds is independent of the number of occurrences of a cue type in one single diagram, while the measure of percentages accounts for the differences in the number of times that cue types occur in one diagram (e.g., percentages of correct boxes are generally higher compared to the other cue types as correct boxes was the diagram cue that occurred most

often in a single diagram). Fig. 4 shows a visualization of the measures.

Furthermore, the number of transitions (i.e., saccades that start in one AOI and end in another AOI) were analysed, distinguishing transitions *within* and *between* diagram types and furthermore between *matching content* and *positions* of the own and standard boxes (from now on *transition types*). For an example, see Fig. 1, where the content of the lower left box in the student's (own) diagram matched the lower box in the middle of the standard, hence matching on content but not position. All transitions were provided as percentages per transition type (see Table 7 for further explanation).

Cued Retrospective Reporting. The eye-movement recordings for the cued retrospective reports were made during the diagram presentation-screens and for the same three out of six trials, plus the practice trial, across participants. Based on Study 1, the trials that resulted in the lowest rate of perfect performance were chosen, so that the chance that students would make errors and use the standard would be higher. The cued retrospective reporting set-up was facilitated with BeGaze (version 3.7), where gaze-replays were presented at 50% of the original speed (cf. Brand-Gruwel et al., 2017) and participants' voices were recorded. Fixations were displayed as blue circles with the size displaying the duration of fixations (50px = 100 ms) which were connected by 2px lines (representing saccades) that faded out after 2 s (see Appendix H). In two instances, the experimental software crashed during the diagram compilation of one of the three trials of interest. In one instance, a trial replay of another trial was compiled for the retrospective reporting part. In the other instance, only two trials (plus the practice trial) were shown for retrospective replay, so a total of 65 trials/retrospective reports was analysed.

Based on van Gog et al. (2005), the instructions were as follows: *This is the practice/first/second/third of three recordings I would like to show you. Please watch it and tell me what you were thinking when you received a correct diagram next to your own diagram. If you would like to pause the recording, press the space bar, and press it again to continue.* When participants fell silent for more than 5 s, they were prompted with *Could you please continue to think aloud?* When the screen-recording of the diagram study phase was finished, the second instruction followed: *This is where you made a judgement about your test score. Could you tell me what you were thinking while making your judgement and what you based your judgement on? You may name several things here.*

After the participant finished with the second question, two questions were asked to be able to estimate the validity of the retrospective reports, namely 1. *On a scale from 1 to 5, how sure are you that these were your own eye-movements?* And 2. *And again, on a scale from 1 to 5, how confident are you that you just explained what you were actually thinking during the experiment, and not just describing what your eye-movements could have meant here?* Both questions were followed up with *Okay, and could you explain why?* if no explanations were given for each of the questions.⁵

3.2.3. Procedure

At the beginning of each session, participants were asked to provide their written informed consent and received a general explanation about the procedure of the study. The participants were then seated alone in front of a desktop-computer with the eye-tracker in a soundproof and windowless room (with no known sources of vibration), with an illuminance of 5.73 cd/m³ which was the same across all sessions. Communication with the experimenter was facilitated via an intercom.

⁵ Analysing the reports showed that participants were (if at all) only surprised and unsure about whether they were reporting on their own eye movements during the practice trial but were relatively sure to see their own eye-movements during the trials that followed the practice and were analysed for our study. The scores indicating how sure participants were to have reported what they were actually thinking did not provide any evidence to suggest they were only describing the eye movements that were displayed to them.

The experiment started with a 9-point machine-controlled calibration (with a slowly pulsating black fixation-cross on a white screen) which was started by the experimenter once the eye-gaze of the participant was visibly stable on the screen. The calibration was followed by a 5-point validation (with the same targets) and then by the drift-check (i.e., the participant was presented with the same target in the centre of the screen and was required to look at the target for 5 s). The calibration was repeated up to four times if validation values were greater than 1.3° of the visual angle. The background of the calibration, validation, and single fixation cross was white, matching the background colour of the rest of the (online) experiment. Identical to Study 1, participants first watched the instruction video and then proceeded through the six trials in randomized order while their eye-movements were recorded. Finally, another drift check was executed.⁶

The time elapsed between calibration and final drift check was approximately 30–40 min. The participant had a two to five-minute break while the experimenter prepared the gaze-recordings for replay and was allowed to leave the headrest. The participant then received instructions for the cued retrospective verbal reporting, which was followed by a practice run where the gaze-recordings of the practice trial were displayed. This was followed by the three trials for which the cued verbal reports were analysed. The total duration of the session was approximately 60 min.

3.2.4. Analyses

Multilevel Analyses. A multilevel model was fitted with the average total fixation duration per diagram box in seconds as dependent variable and diagram type (own and standard diagram), cue types (correct, commission, and omission), and their interaction as predictors. The nested structure accounted for by the model was cue and diagram types nested within participants. The same model parameters as in Study 1 were provided and *p* values were Tukey-adjusted for multiple pairwise comparisons.

Coding the Think-Aloud Protocols. The think-aloud protocols were coded in two rounds. First, 20% of the protocols were segmented by the first author and a colleague. A segment refers to a (part of a) sentence that is meaningful in itself, independently of coding categories. Since the overlap between raters ranged from 83.16% to 88.64%, the interrater-reliability was good (cf. Strijbos et al., 2006). The first author subsequently segmented the remaining protocols. In a second step, the segments were coded with regard to cue types (i.e., whether students named a particular cue, such as *I filled this in correctly* for cue correct boxes), diagram comparisons both related to content (e.g., *I worded it differently*) and processing (e.g., *Then I went back and compared*), the diagram type, contextual information, and JOLs (see Table 4 for all codes, definitions, and examples). To establish the interrater reliability of this coding scheme, two raters (the first author and a colleague), coded 20% of the segmented protocols, resulting in $\kappa = 0.834$ (agreement = 85.10%). Segments during the diagram study screen and during the JOL screen are reported both combined (to establish a comprehensive picture of students' awareness of cues and diagram aspects), and separately for the JOL screen to be able to distinguish which cues students mentioned related to their JOLs.

Comparative Results of Performance Data Across Study 1 and 2. Next to the descriptive statistics of JOLs, monitoring accuracy, and text comprehension, that were provided in Table 1, for comparison with the descriptive statistics of Study 1, cue values, cue diagnosticity and cue utilization of both studies were computed and reported in Table 9 for a comprehensive picture of the comparability of both studies.

⁶ The analysis of potential drift was not included in this article.

Table 6
(Selection of) Pairwise Comparisons of Total Fixation Duration per Diagram and Cue Type.

Comparisons of Combinations Cue per Diagram Type			<i>B</i>	<i>S.E.</i>	<i>t</i>	<i>p</i>	<i>b</i>
Within Own Diagram Comparisons							
Commission	-	Correct	-0.13	0.25	-0.52	1.000	-0.07
Commission	-	Given Box	0.71	0.32	2.20	.354	0.38
Correct	-	Given Box	0.84	0.26	3.27	.025	0.45
Within Standard Diagram Comparisons							
Commission	-	Correct	1.58	0.24	6.61	<.001	0.84
Commission	-	Omission	1.41	0.41	3.47	.013	0.75
Commission	-	Given Box	3.11	0.28	11.17	<.001	1.65
Correct	-	Omission	-0.17	0.36	-0.48	1.000	-0.09
Correct	-	Given Box	1.52	0.20	7.53	<.001	0.81
Omission	-	Given Box	1.70	0.39	4.40	<.001	0.90
Between Standard and Own Diagram Comparisons							
(Standard) Commission	-	(Own) Commission	3.39	0.31	11.10	<.001	1.80
(Standard) Correct	-	(Own) Correct	1.68	0.15	11.57	<.001	0.89

Note. *Standard* refers to the correct standard. *(Standard) Commission*, for example, refers to correct boxes in the standard that correspond to a commission error in the own diagram. P-values were adjusted with Tukey method for comparing a family of eight estimates.

Table 7
Number of Transitions per Trial and Transition Type.

Label of Type	Transition Type	<i>n</i>	% of transitions relative to total number of all transitions	% of transitions relative to number of transition type	
				%	(Of transition type)
1.	Total number of all transitions	2426	100		
	Transitions between own diagram and standard	963	39.69	39.69	(all transitions)
1.a	Correct box in own diagram – corresponding box in standard	720	29.68	74.77	(type 1)
	Matching Position	115	4.74	15.97	(type 1.a)
	Matching Content	66	2.72	9.17	
	Both Matching	319	13.15	44.31	
	Neither Matching Position nor Content	220	9.07	30.56	
1.b	Error in own diagram – corresponding box in standard	186	7.67		(type 1)
1.b.i	Commission – Standard	169	6.97	90.86	(type 1.b)
	Matching Position	84	3.46	49.70	(type 1.b.i)
	Matching Content	5	0.21	2.96	
	Both Matching	28	1.15	16.57	
	Neither Matching Position nor Content	52	2.14	30.77	
1.b.ii	Omission - Standard	17	0.70	9.14	(type 1.b)
	Matching Position*	7	0.29	41.18	(type 1.b.ii)
	Not Matching Position	10	0.41	58.82	
1.c	Given box in own diagram and standard	57	2.35	5.92	(type 1)
2.	Transitions within diagrams	1463	60.31	60.31	(all transitions)
2.a	Within Own Diagram	317	13.07	21.67	(type 2)
2.b	Within Standard	1146	47.24	78.33	

Note. A Chi-square test comparing the number of transitions *within* the own diagram ($n = 317$) and the standard diagram ($n = 1146$) showed it was significantly lower in the own diagram, $\chi^2(1) = 469.75, p < .001$. The number of transitions *between* the two diagrams ($n = 963$) was significantly higher than the number of transitions within the own diagram, $\chi^2(1) = 326.03, p < .001$, and significantly lower than the number of transitions within the standard diagram, $\chi^2(1) = 15.88, p < .001$.

4. Results

4.1. Judgements of learning, monitoring accuracy, and text comprehension

See Table 1 for the descriptive statistics for monitoring accuracy, test scores, JOL₁, and JOL₂ of Study 2.

4.2. Processing of the standard (RQ 2.1)

Fixation Durations. See Table 5 for descriptive statistics related to the average total fixation duration per box and percentage total fixation duration per box. There was a difference between how many boxes were

fixated in the standards, compared to how many boxes were fixated in the own diagrams. While students fixated almost all boxes in the standard (e.g., 115 of 117 given boxes in the standards), many boxes in the own diagrams were not fixated (e.g., 65 of 117 given boxes). The average total fixation duration on a single box (or AOI) per cue type showed that across all correct boxes of all participants, one correct box in the own diagram was on average fixated 1.29 s, while a box in the standard corresponding to such a correct box was on average fixated for 2.93 s. Note that the number of words per own diagram box was larger than the number of words in the standard diagram boxes: own responses varied between 0–21 ($M = 7.86, SD = 3.93$), standard responses varied between 3–12 ($M = 6.13, SD = 2.35$) words. Participants wrote on average 1.73 words more in a box of their own diagram compared to a

Table 8
Average Frequencies and Proportion of Code Categories per Participant and Entire Sample.

Coding Category Types	During Diagram Study & JOL Screen						During JOL Screen Only					
	Segments per coding category		Frequency per participant		Mentioned by:		Segments per coding category		Frequency per participant		Mentioned by:	
	n	%	Mean	SD	n	%	n	%	Mean	SD	n	%
Cue Type Identification^{*2}												
Correct Box ^{*1}	50	3.69	2.94	2.86	17	80.95	8	1.90	1.33	0.52	6	28.57
Correct Relation	2	0.15	1.00	0.00	2	9.52	1	0.24	1.00	–	1	4.76
Incorrect Box Commission	44	3.25	2.75	1.65	16	76.19	7	1.66	1.17	0.41	6	28.57
Incorrect Box Omission	8	0.59	1.60	0.89	5	23.81	3	0.71	1.50	0.71	2	9.52
Incorrect, Not-Specified ^{*1}	34	2.51	2.00	1.06	17	80.95	5	1.18	1.00	0.00	5	23.81
Incorrect Relation	5	0.37	1.25	0.50	4	19.05	–	–	–	–	–	–
Content-Related Diagram Comparisons^{*1}												
Mismatch-Identification	45	3.32	2.65	1.41	17	80.95	10	2.37	1.43	0.53	7	33.33
Match-Identification	36	2.66	1.89	1.05	19	90.48	15	3.55	1.50	0.71	10	47.62
Processing-Related Statements												
Own + Standard ^{*1}	65	4.80	3.42	2.59	19	90.48	2	0.47	1.00	0.00	2	9.52
Own	26	1.92	1.86	1.29	14	66.67	1	0.24	1.00	–	1	4.76
Standard	109	8.05	5.45	3.43	20	95.24	9	2.13	1.29	0.49	7	33.33
Other/General	90	6.65	5.00	4.84	18	85.71	1	0.24	1.00	–	1	4.76
Standard Diagram												
Test Preparation ^{*1}	105	7.75	5.53	4.22	19	90.48	10	2.37	2.00	1.00	5	23.81
(Meta-) Comments	118	8.71	5.62	4.73	21	100.0	33	7.82	3.00	2.79	11	52.38
Own Diagram^{*1}												
(Meta-) Comments	54	3.99	3.60	3.54	15	71.43	10	2.37	1.67	0.82	6	28.57
Context-Related												
Previous Trials	19	1.40	2.11	1.54	9	42.86	16	3.79	2.00	1.69	8	38.10
Topic/Text-Related	32	2.36	3.20	1.81	10	47.62	17	4.03	2.12	1.25	8	38.10
Judgements of Learning												
Metacognitive judgements	249	18.39	11.86	4.63	21	100.0	155	36.73	7.38	2.97	21	100
Judgements of Learning	76	5.61	4.00	1.73	19	90.48	73	17.30	3.84	1.89	19	90.48
Irrelevant Codes for Analysis												
Emotions	4	0.30	2.00	1.41	2	9.52	0	0	–	–	0	0
Other	183	13.52	9.15	6.34	20	95.24	46	10.90	2.71	1.57	17	80.95

Note. There were a total of 1354 segments within the think-aloud protocols of 21 participants, 422 of those segments were voiced during the JOL screen. *Frequency per participant* refers to the average number of utterances per code per participant across our sample, *Mentioned by* refers to the total number of participants that mentioned the code. *During JOL Only* is the summary of codes mentioned during the JOL screen. ^{*1}refers to codes interpreted for RQ1, ^{*2} to codes interpreted for RQ2.

Table 9
Descriptive Statistics Cue Values, Cue Diagnosticity and Cue Utilization.

Cue-	Task	Correct			Commission			Omission		
		M (SD)	[Min, Max]	NA	M (SD)	[Min, Max]	NA	M (SD)	[Min, Max]	NA
Value	DO ₁	3.12 (0.88)	[1, 4]	0	0.64 (0.73)	[0, 3]	0	0.28 (0.60)	[0, 3]	0
	DS ₁	2.97 (0.97)	[0, 4]	0	0.73 (0.77)	[0, 3]	0	0.36 (0.65)	[0, 3]	0
	DS ₂	3.02 (0.95)	[1, 4]	0	0.82 (0.91)	[0, 3]	0	0.26 (0.55)	[0, 3]	0
Diagnosticity	DO ₁	0.60 (0.29)	[−0.11, 1.00]	0	−0.34 (0.41)	[−0.91, 0.71]	0	−0.51 (0.35)	[−1.00, 0.17]	4
	DS ₁	0.28 (0.41)	[−0.37, 0.87]	3	−0.21 (0.57)	[−1.00, 0.61]	5	−0.19 (0.43)	[−0.77, 0.46]	4
	DS ₂	0.29 (0.49)	[−0.46, 0.88]	7	−0.22 (0.57)	[−0.85, 0.54]	7	−0.11 (0.44)	[−0.76, 0.45]	12
Utilization	DO ₁	0.45 (0.37)	[−0.72, 0.77]	0	−0.14 (0.39)	[−0.85, 0.43]	0	−0.52 (0.46)	[−1.00, 0.79]	4
	DS ₁	0.44 (0.32)	[−0.11, 0.97]	1	−0.17 (0.38)	[−0.86, 0.45]	2	−0.33 (0.36)	[−0.89, 0.45]	3
	DS ₂	0.56 (0.27)	[−0.03, 0.88]	2	−0.38 (0.38)	[−0.88, 0.49]	1	−0.48 (0.32)	[−0.91, 0.00]	7

Note. The diagram *Cue Values* indicate the average number of occurrences of *correctly* completed boxes, *commission* errors, and *omission* errors per diagram. Descriptive statistics are reported for the Diagramming + Standard (*task*) of Study 1 (*DS₁*) and 2 (*DS₂*), and for Diagramming-Only (*DO₁*) task in Study 1. Pearson correlation coefficients are provided for *Cue Diagnosticity* and *Cue Utilization* and reflect correlations between Cue Values and Test scores, and Cue Values and JOL₂ scores, respectively. *NA* indicate excluded participants per condition due to standard deviations of 0 across six trials.

box of the standard. This makes the longer fixations on the standard stand out even more. Furthermore, average percentages of total fixation duration across participants (see last column in Table 5) show that participants fixated the standard on average for 61.42% of the duration of the diagram study screen, which is on average almost three times as long as compared to their own diagram (21.59%).

ICCs indicated that 10% of the variance was explained by the grouping structure (multiple trials per participant). There were significant effects of diagram type, $B = 3.39, S.E. = 0.31, CI = [2.79 - 3.99], p < .001$, cue type, $B = -0.71, S.E. = 0.32, CI = [-1.34 - -0.08], p = .028$, and their interaction, $B = -1.71, S.E. = 0.34, CI = [-2.37 - -1.05], p$

$< .001$, for predicting fixation duration.

Post-hoc pairwise comparisons showed that total fixation durations of boxes in students' own diagrams did not significantly differ based on their cue types. That is, there were no significant differences in total fixation durations between commission errors, correctly completed, given and empty (omission) boxes within students' own diagrams (see Table 6). However, boxes in the standard corresponding to commission errors were on average fixated significantly longer than boxes corresponding to all other cue types (i.e., correct, omission, or given) in the standard. Furthermore, boxes in the standards were generally fixated significantly longer than their corresponding cue type boxes in students'

own diagram, with the exception of given boxes, for which total fixation durations did not significantly differ between the diagram types (see [Appendix I](#) for all pairwise comparisons).

Number of Transitions between AOIs. There were between 55 and 204 transitions per participant ($M = 121.30$, $SD = 38.00$). Most fixation transitions occurred within diagrams (60.31%), and especially within in the standard (78.33% of the within diagram transitions), see [Table 7](#) for all numbers and percentages of transition types. Most transitions between diagrams occurred between correctly completed boxes in the own diagram and boxes in the standard that matched both in position and content (44.31%).

Cued Retrospective Reporting Related to Processing of the Standard. See [Table 8](#) for the average frequencies and proportion of code categories per trial and total mentions of all coding categories. Most of the students mentioned comparing the own and the standard at least once (90.48%), with an average of 3.42 references per participant. Over 90% said to have used the standard as preparation for the test, with an average of 5.53 references per participant. Over 80% of the students made general comments about matching (90.84%; 1.89 references per participant on average) and mismatching (80.95%; average per participant 2.65 references) attributes of the standard and their own diagram.

4.3. Inferred and used diagram cues for monitoring judgements (RQ2)

Retrospective Reporting Related to Cue Use. Regarding inferred diagram cues, we found that most students were able to identify diagram cues during the retrospective reporting. More specifically, [Table 8](#) shows that over 80% of the students identified at least one correctly completed box in their own diagrams (with an average of 2.94 references per participant), and also over 80% of the students found at least one incorrectly completed box in their own diagrams, while not specifying whether it was an omission or commission error. 76.19% of the students specifically mentioned commission errors and 23.81% mentioned omission errors. Correct or incorrect relations were only mentioned by 9.52% and 19.05% of the students, respectively.

During the JOL screen, 28.57% of the students mentioned correctly identified boxes as basis for their JOLs and 47.62% mentioned matching attributes of the standard and their own diagram. Again, 28.57% of the students mentioned identified commission errors as basis for their JOLs and over 50% made (meta-)comments about the standard while answering the question on which information they based their JOL. Over 38% made comments about the text or topic and compared the current trial with previous trials when making the JOL.

4.4. Cue values, cue diagnosticity and cue utilization across studies

See [Table 9](#) for the descriptive statistics related to the value (i.e., the number of occurrences of a cue), diagnosticity and utilization of each Cue Type. None of the cue values, cue diagnosticity or cue utilization values significantly differed between the matching conditions of Study 1 and 2 (DDS). Furthermore, all cue values, cue diagnosticity or cue utilization coefficients did not significantly differ between the two conditions of Study 1.

4.5. Discussion Study 2

The results of Study 2 showed that students looked at the standard on average more than twice as long compared to their own diagram, and fixated boxes in the standard that corresponded to their commission errors the longest. Transitions of fixations between the standard and the own diagram occurred most often between matching content and position of correct diagram boxes. Most students reported correctly and incorrectly completed boxes during the cued retrospective reporting, voiced to have compared the two diagrams, and reported to have used the standard as preparation for the comprehension test of each trial. Thus, these findings provide insight into how the beneficial effects of

receiving a standard after diagramming on monitoring accuracy and text comprehension that were found in Study 1, came about.

5. General discussion

5.1. Effects of diagramming and receiving a standard on monitoring

In contrast to our expectation, there was no significant interaction effect between diagramming and receiving a standard. However, we did find a main effect of receiving a standard and exploratory follow-up analyses showed that this effect was mainly driven by the combined Diagramming + Standard condition. This was the only condition to show significantly better monitoring accuracy compared to the No-Diagram control condition.

The fact that we did find, with our immediate design, a significant effect of the combination of diagramming and receiving a correct diagram standard on monitoring accuracy, is very interesting (also from an educational perspective). It suggests that the standard can remedy a drawback that immediate generative activities may have, namely that irrelevant factual (or surface level) information is still available in working memory when completing the activity, and thus, the cues to be gained are less diagnostic for final test performance (during which students have to rely on their situation model of the text; [Griffin et al., 2008](#); [Thiede et al., 2010](#)). By giving students a standard to which they can compare their own answer, in contrast, they get highly diagnostic cues, as they can spot their commission errors and recall of irrelevant information, which improves their monitoring accuracy.

Indeed, the findings from Study 2 suggest that students actively compared the standards to their own answers to gain this kind of information. The eye-movement data showed that students fixated on the standard (on average) almost three times longer than on their own diagram and fixated on boxes in the diagram-standard corresponding to their own commission errors significantly longer than on boxes corresponding to their own correctly completed boxes. The extensive processing of the standard was further underlined by the frequently observed transitions between information in the standard and the own diagram. The verbal data showed that over half of the participants mentioned the standard when asked on which information they based their JOLs. Students also mentioned matching attributes of the own diagram with the standard (i.e., correctly completed boxes) as basis for their JOLs, which makes sense given the correctness of their diagrams. Interestingly, students not only mentioned making use of diagram-based performance cues, but also of contextual cues such as text properties (e.g., difficult text with a lot of information) or topic-related cues (e.g., interest or familiarity with the text topic). This is in line with research of [Dinsmore and Parkinson \(2013\)](#) and [List and Alexander \(2015\)](#) who asked students to self-report their cue utilization and found that most students named text-based justifications (for a more elaborate overview see also [Hacker & Bol, 2019](#)). Most of the participants mentioned correct and incorrect boxes during cued retrospective reporting in Study 2, although they typically did not specify whether incorrect boxes meant omission or commission errors. It is likely, however, that the standard did help them spot the commission errors, given the eye movement data, and this could explain the improved monitoring accuracy which was observed in Study 1.

That we did not find improved monitoring accuracy when only a standard was provided, is in line with the study by [Redford et al. \(2012\)](#), on concept maps. The finding that diagramming only did not significantly improve monitoring accuracy compared to the control condition, is in line with the findings from the immediate diagramming condition from the study by [van Loon et al. \(2014\)](#). Interestingly, JOLs were not adapted after diagramming only. One potential reason for this observation could have been the low number of omission errors per trial, which are easy to spot, in combination with unrecognized commission errors, which are harder to spot without feedback. In other words, the cues students gained from diagramming only in our study may not have

been potent enough to improve their diagnostic cue use while making JOLs and thereby, their monitoring accuracy. Because commission errors during diagramming seem to have a negative effect on monitoring accuracy (cf. van De Pol et al., 2020, who report that students with low monitoring accuracy made more commission errors while completing equally many diagram boxes compared to students with high monitoring accuracy), this finding again highlights the usefulness of providing students with the means (such as a standard) to identify commission errors made during the diagramming task.

5.2. Effects of diagramming and receiving a standard on text comprehension

We found that completing diagrams and then receiving a correct diagram as standard, led to significantly higher text comprehension compared to the other experimental conditions. Diagramming only or receiving a standard only did not lead to significantly better comprehension than in the control condition. Interestingly, the data regarding the change in JOL ratings did suggest that participants in both standard conditions felt that the standard increased their comprehension. That is, they made significantly higher JOLs after studying the standard than after reading the text, while participants who did not get standards did not significantly change their JOLs. However, only the Diagramming + Standard condition showed significantly better text comprehension (and monitoring accuracy) compared to the other conditions.

The finding that the Standard-Only condition did not show better text comprehension is interesting and important as it also shows that the effect of the combined condition was not merely due to the fact that those students had access to more information than students who did not get the standard. That is, the standard basically provided a restudy opportunity of the most important information, but if this was the main driver of the beneficial effect on text comprehension, one would also expect such benefits in the Standard-Only condition. Yet, students first needed to engage in the diagramming task, in order to benefit from the additional information provided by the standards. This can be either due to experiential cues, that is, experiencing that the diagramming activity is hard, which may lead them to process the standard more thoroughly than students who only received the standard, or to the fact that the standard allows for correcting misunderstandings and repairing knowledge gaps, or both. In any case, the findings from Study 2 (which had only the Diagram + Standard condition) also showed that the standard was extensively processed by learners, which may explain the beneficial effects on text comprehension.

Note though, that the finding that a standard only did not improve text comprehension, is at odds with the findings of McCrudden et al. (2007), who found that simply showing causal diagrams depicting relations of texts (without previous self-diagramming) already improved text comprehension (possibly, the fact that their text was substantially longer than our texts, played a role). Moreover, one would also have expected immediate diagramming only to benefit text comprehension compared to the control condition, as prior studies without standards showed that immediate generative activities may not foster monitoring accuracy (when feedback is not available), but do foster text comprehension (e.g., Anderson & Thiede, 2008; van Loon et al., 2014). However, this was not the case in our study.

5.3. Limitations and future directions

A limitation of Study 1 was our rather proficient sample in combination with a relatively small response scale for JOLs, text comprehension, and monitoring accuracy, ranging from 0 to 4. There was little opportunity for participants to overestimate their comprehension, which is usually reported to be a main reason for poor monitoring accuracy (Hacker & Bol, 2019). Yet, we did find a significant improvement in monitoring accuracy after diagramming and receiving a standard. However, this improvement was probably not due to reduced

overestimation, as JOLs actually increased after studying the standard; but so did students' test scores (at least in the diagramming + standard condition). Thus, the improved accuracy seems to have been mainly related to an alignment of JOL and test scores, rather than to lowering JOLs after overestimations. Hence, future research should attempt to replicate our findings with a less proficient sample (e.g., secondary school students), to see whether providing standards after immediate diagramming would also reliably reduce overestimations of JOLs.

A potential limitation of Study 2 is that we cannot rule out that students' memory of their own diagram could be a possible alternative explanation for why viewing times on their own diagram were shorter. However, even if most information about the own diagram was retained in memory while comparing their own diagram to a standard, the findings that the standards were indeed studied and used to identify mistakes and prepare for the test still provide valuable insights to answering our research questions. Future research could consider relating the gaze data with the retrospective reports more systematically, and to relate this process data in turn to the monitoring accuracy and text comprehension scores in a bigger sample, to further unravel students' awareness and use of available cues for their monitoring judgements.

Furthermore, it seems as if participants did not think of the combinations of completed diagram boxes, in terms of correct or incorrect relations between the different boxes (only about the correctness or incorrectness of individual boxes), whereas the relations are an important part of text comprehension. However, this finding might be an artefact of our segmentation and coding scheme. For example, protocol-segments such as *I tried to remember the sequence of these two boxes* were coded as processing of the standard as preparation for the test, while the participant may have done so because they inferred that they did not have that sequence, or the correct relation, in their own diagram. Yet it might also mean that when using standards and giving students explicit comparison or self-assessment instructions, it might be worthwhile to also draw their attention explicitly on the relations. Furthermore, all participants mentioned content of the standard and while doing so (especially when naming specific boxes), it is likely that they were aware of and used cue values, such as whether the boxes were correct or incorrect. However, when they did not specifically mention the cue values or cue utilization for the monitoring judgements, we could not code it as such.

Finally, in a less proficient sample, it would also seem interesting to investigate whether students who make less accurate monitoring judgements adopt less efficient ways of processing correct diagram-standards. For example, by spending less time reading and fixating on the standard or making fewer comparisons between matching positions or content of diagram boxes. Identifying processing differences between students with high versus low monitoring accuracy during their standard study could then help to further advance interventions that aim to improve monitoring accuracy.

6. Conclusion and educational implications

Our findings showed that having students complete a causal diagram immediately after reading a text and then study a correct diagram-standard improved both monitoring accuracy and text comprehension. Eye movement and verbal protocol data showed that this effect likely resulted from the fact that participants studied the provided standards, identified mistakes they made in their own diagram, and (thereby) prepared for the upcoming comprehension test. In contrast to most prior studies, we used an immediate design, which seems to be more practical in real classroom settings. Our results therefore suggest that it might be helpful for (university) students to get access to a correct standard after completing a generative learning task, such as causal diagramming. Building upon our results related to causal diagrams, modern tools such as automatically generating concepts maps of written texts (Lachner et al., 2017) might hold promise for facilitating the creation of diagram

standards without further burdening teachers with the task of preparing standards for the texts to be studied. Last but not least, our findings are promising for (digital) classroom interventions aimed at improving students' monitoring accuracy and text comprehension.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

A link to our R analysis and data cleaning scripts, together with

artificial data that imitate our real data, is provided in Appendix D.

Acknowledgements

The authors would like to thank Mirjam Moerbeek for her support with the power analyses for multilevel models, Anne Milder for transcribing the audio data, and Jonne Bloem, Jael Draijer, and Marloes van Dijk for helping to establish interrater reliabilities of the different coding schemes. This research was funded by a grant from the Netherlands Initiative for Education Research (project number 40.5.18300.024).

Appendix A. Additional translated instructions (in Blue) of Introduction video for Diagramming + Standard condition

Text screen (Presented for all conditions).

Each trial starts with the task of reading a text

First JOL Screen (Presented for all conditions).

Each trial ends with a question that tests your text comprehension. After reading each text, we ask you to make an judgement about...

The test consists of a question that you should answer with four possible relations from the text. For each correct relation you get one point, so you may now choose a score between 0 and 4.

Diagram Completion Screen (Diagramming-Only and Diagramming + Standard Conditon).

Next, we ask you to fill in four empty boxes in a diagram about the text

Diagram Study Screen (Diagramming + Standard Conditon).

*Then you will be shown another (correct) diagram next to your own
Use the diagrams to prepare for the test*

Second JOL Screen (Presented for all conditions).

Then we will ask you again to estimate your score on the test question

See [Appendix A](#) for all other instructions that are displayed both during the video and the experimental trials.

Appendix B. Two translated sample texts

The Suez Canal



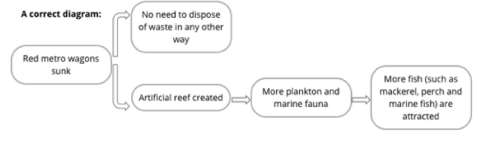
The Suez Canal, which connects the Indian Ocean and the Mediterranean Sea with each other, is of great importance to the world. Originally, there was no natural water connection between the Atlantic and the Indian Ocean. Between these two seas is a desert. This meant that trading ships that travelled from the harbour 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 harbour city of Jeddah to the harbour 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.

Botox

Botox is the abbreviation of Botulinium Toxin, this is a poison that is produced by the bacterium Clostridium botulinum. 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.

The remaining (Dutch) texts are available upon request.

Appendix C. Example of a translated trial

<p>Sunken subway cars</p> <p>The majority of the ocean floor is around 2 to 2.5 miles deep and has relatively little change in elevation. This is called the abyssal plain. In the Atlantic Ocean off the East Coast of the United States, 700 train cars from the New York subway have been submerged. The old, red New York subway cars are also referred to as redbirds. By dumping these train cars in the ocean, they no longer had to be disposed of in other ways. At the same time, sinking these train cars created an artificial reef in the abyssal plain. Since the sinking of the first train cars in 2001, this new artificial reef has resulted in the amount of plankton and marine fauna increasing by 400 times. The sandy seabed of the East Coast of the United States normally does not provide as much habitat for marine animals. This increase in plankton and marine fauna around the artificial reef is now attracting more mackerel, bass, and other saltwater fish.</p> <p style="text-align: right;">Next</p>	<p>How many points, do you think, would you score on a test about the text Metrowagons sunk ?</p> <div style="display: flex; justify-content: space-around; align-items: center;"> <div style="border: 1px solid gray; padding: 5px; width: 40px; height: 40px; display: flex; align-items: center; justify-content: center;">0</div> <div style="border: 1px solid gray; padding: 5px; width: 40px; height: 40px; display: flex; align-items: center; justify-content: center;">1</div> <div style="border: 1px solid gray; padding: 5px; width: 40px; height: 40px; display: flex; align-items: center; justify-content: center;">2</div> <div style="border: 1px solid gray; padding: 5px; width: 40px; height: 40px; display: flex; align-items: center; justify-content: center;">3</div> <div style="border: 1px solid gray; padding: 5px; width: 40px; height: 40px; display: flex; align-items: center; justify-content: center;">4</div> </div>
<p>Please complete the diagram for the text Sunken subway cars. If you cannot complete a box, type in a ?</p>  <p style="text-align: right;">next</p>	<p>See your own and a correct diagram below.</p> <p>Your diagram:</p>  <p>A correct diagram:</p>  <p style="text-align: right;">next</p>
<p>How many points do you think you could score on a test about the text now?</p> <div style="display: flex; justify-content: space-around; align-items: center;"> <div style="border: 1px solid gray; padding: 5px; width: 40px; height: 40px; display: flex; align-items: center; justify-content: center;">0</div> <div style="border: 1px solid gray; padding: 5px; width: 40px; height: 40px; display: flex; align-items: center; justify-content: center;">1</div> <div style="border: 1px solid gray; padding: 5px; width: 40px; height: 40px; display: flex; align-items: center; justify-content: center;">2</div> <div style="border: 1px solid gray; padding: 5px; width: 40px; height: 40px; display: flex; align-items: center; justify-content: center;">3</div> <div style="border: 1px solid gray; padding: 5px; width: 40px; height: 40px; display: flex; align-items: center; justify-content: center;">4</div> </div>	<p>On the next screen, the test starts over the text Sunken subway cars.</p> <p style="text-align: center;">Are you ready?</p> <p style="text-align: right;">next</p>
<p>Metro wagons have been sunk in the United States. The sinking of the metro wagons has several consequences. What are these consequences?</p> <p>Try to give as complete an answer as possible, mentioning four connections. Please fill in a ? if you don't know anything.</p> <p style="text-align: center;">Good luck!</p> <div style="border: 1px solid gray; padding: 10px; min-height: 50px;"> <p>Naming four relations here.</p> </div> <p style="text-align: right;">Next</p>	

Appendix D

The R scripts for data processing and analysis of both studies are available (under a MIT licence) on Github (https://github.com/SopBra/NRO-PROO_Article_1_Effects_Of_Feedback_Standards_During_Diagramming) and Zenodo (<http://doi.org/10.5281/zenodo.10204078>). The scripts on Github will be maintained and potentially updated in future based on feedback and improved R functions. Scripts on Zenodo will reflect a reproducible snapshot of the data processing at the moment of publication of this article.

SopBra (2023). SopBra/NRO-PROO_Article_1_Effects_Of_Feedback_Standards_During_Diagramming: First Release. <https://doi.org/10.5281/zenodo.10204078>

Appendix E. Additional Exploratory Analyses: Pairwise Comparisons of Conditions

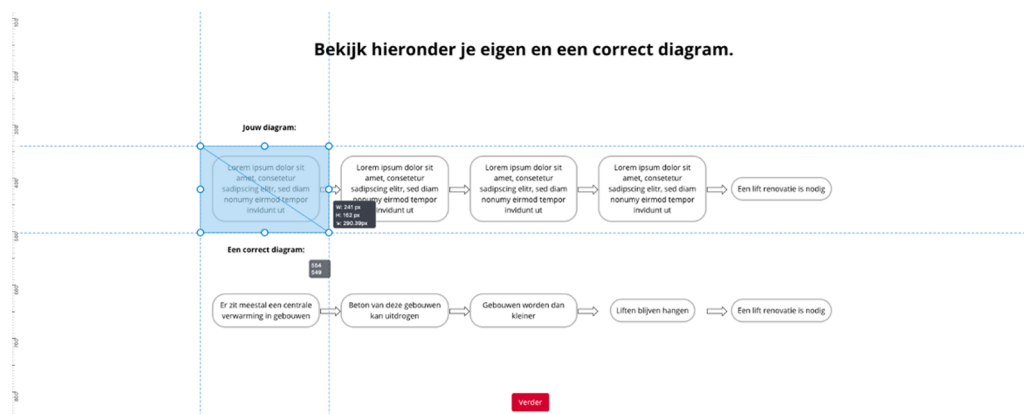
	B	S.E.	df	t	p	b
Monitoring Accuracy ~						
DO – DS	0.30	0.12	74.37	2.54	.061	0.41
DS – SO	-0.07	0.12	75.08	-0.58	.938	-0.09
ND – DS	0.33	0.12	74.78	2.80	.032*	0.46
DO – SO	0.23	0.11	75.10	2.01	.192	0.32
ND – DO	0.04	0.12	74.79	0.30	.990	0.05
ND – SO	0.27	0.12	75.51	2.29	.110	0.37
Text Comprehension ~						
DO – DS	-0.57	0.15	74.52	-3.85	.001*	-0.73
DS – SO	0.43	0.15	75.07	2.89	.025*	0.54
ND – DS	-0.75	0.15	74.84	-4.96	<.001*	-0.95
DO – SO	-0.15	0.15	75.08	-1.01	.746	-0.19
ND – DO	-0.18	0.15	74.85	-1.18	.639	-0.22
ND – SO	-0.32	0.15	75.40	-2.19	.136	-0.41

Appendix F. All pairwise comparisons of JOL-Instance and condition

Contrast	B	S.E.	df	t	p	b
JoL1 No-Diagram – JoL2 No-Diagram	0.17	0.10	865.00	1.70	.687	0.23
JoL1 No-Diagram – JoL1 Diagramming-Only	0.05	0.19	97.61	0.28	1.000	0.07
JoL1 No-Diagram – JoL2 Diagramming-Only	0.00	0.19	97.61	-0.02	1.000	-0.01
JoL1 No-Diagram – JoL1 Diagramming + Standard	0.30	0.20	97.61	1.52	.797	0.40
JoL1 No-Diagram – JoL2 Diagramming + Standard	-0.61	0.20	97.61	-3.08	.053	-0.82
JoL1 No-Diagram – JoL1 Standard-Only	0.23	0.19	97.61	1.19	.932	0.31
JoL1 No-Diagram – JoL2 Standard-Only	-0.37	0.19	97.61	-1.91	.549	-0.49
JoL2 No-Diagram – JoL1 Diagramming-Only	-0.11	0.19	97.61	-0.58	.999	-0.15
JoL2 No-Diagram – JoL2 Diagramming-Only	-0.17	0.19	97.61	-0.88	.987	-0.23
JoL2 No-Diagram – JoL1 Diagramming + Standard	0.13	0.20	97.61	0.67	.998	0.18
JoL2 No-Diagram – JoL2 Diagramming + Standard	-0.77	0.20	97.61	-3.92	.004	-1.04
JoL2 No-Diagram – JoL1 Standard-Only	0.06	0.19	97.61	0.32	1.000	0.08
JoL2 No-Diagram – JoL2 Standard-Only	-0.53	0.19	97.61	-2.78	.113	-0.72
JoL1 Diagramming-Only – JoL2 Diagramming-Only	-0.06	0.10	865.00	-0.61	.999	-0.08
JoL1 Diagramming-Only – JoL1 Diagramming + Standard	0.24	0.19	97.61	1.26	.911	0.33
JoL1 Diagramming-Only – JoL2 Diagramming + Standard	-0.66	0.19	97.61	-3.39	0.022	-0.89
JoL1 Diagramming-Only – JoL1 Standard-Only	0.18	0.19	97.61	0.93	0.983	0.24
JoL1 Diagramming-Only – JoL2 Standard-Only	-0.42	0.19	97.61	-2.22	0.351	-0.57
JoL2 Diagramming-Only – JoL1 Diagramming + Standard	0.30	0.19	97.61	1.56	0.772	0.41
JoL2 Diagramming-Only – JoL2 Diagramming + Standard	-0.60	0.19	97.61	-3.09	0.051	-0.81
JoL2 Diagramming-Only – JoL1 Standard-Only	0.23	0.19	97.61	1.23	0.920	0.32
JoL2 Diagramming-Only – JoL2 Standard-Only	-0.36	0.19	97.61	-1.91	0.549	-0.49
JoL1 Diagramming + Standard – JoL2 Diagramming + Standard	-0.90	0.10	865.00	-9.22	<0.001	-1.22
JoL1 Diagramming + Standard – JoL1 Standard-Only	-0.07	0.19	97.61	-0.36	1.000	-0.09
JoL1 Diagramming + Standard – JoL2 Standard-Only	-0.66	0.19	97.61	-3.46	0.017	-0.90
JoL2 Diagramming + Standard – JoL1 Standard-Only	0.83	0.19	97.61	4.35	0.001	1.13
JoL2 Diagramming + Standard – JoL2 Standard-Only	0.24	0.19	97.61	1.24	0.916	0.32
JoL1 Standard-Only – JoL2 Standard-Only	-0.60	0.09	865.00	-6.38	<0.001	-0.80

Note. JOL1: First Judgement of Learning, JOL2: Second Judgement of Learning. P-values were Tukey-corrected for multiple comparisons.

Appendix G. Example AOI definition in own diagram



Appendix H. Screenshot of gaze pattern as displayed for cued retrospective verbal reporting



Appendix I. All pairwise comparisons between fixations of combinations diagram type - cues type

	<i>b</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>	<i>b</i>
<i>Within Own Diagram</i>						
Own Commission - Own Correct	-0.13	0.25	1041.99	-0.52	1.000	-0.07
Own Commission - Own Given	0.71	0.32	1035.38	2.20	.354	0.38
Own Commission - Own Omission	0.65	0.47	1037.18	1.39	.863	0.35
Own Correct - Own Given	0.84	0.26	1033.25	3.27	.025	0.45
Own Correct - Own Omission	0.78	0.43	1036.13	1.83	.599	0.42
Own Given - Own Omission	-0.06	0.48	1035.15	-0.12	1.000	-0.03
<i>Within Standard Diagram</i>						
Standard Commission - Standard Correct	1.58	0.24	1042.02	6.61	<.001	0.84
Standard Commission - Standard Given	3.11	0.28	1037.06	11.17	<.001	1.65
Standard Commission - Standard Omission	1.41	0.41	1041.52	3.47	.013	0.75
Standard Correct - Standard Given	1.52	0.20	1030.77	7.53	<.001	0.81
Standard Correct - Standard Omission	-0.17	0.36	1039.71	-0.48	1.000	-0.09
Standard Given - Standard Omission	-1.70	0.39	1037.77	-4.40	<.001	-0.90
<i>Between Own and Standard Diagram</i>						
Own Commission - Standard Commission	-3.39	0.31	1030.49	-11.10	<.001	-1.80
Own Commission - Standard Correct	-1.81	0.25	1041.98	-7.33	.000	-0.96
Own Commission - Standard Given	-0.29	0.28	1037.14	-1.01	.973	-0.15
Own Commission - Standard Omission	-1.98	0.41	1041.47	-4.82	<.001	-1.05
Standard Commission - Own Correct	3.26	0.24	1042.46	13.52	<.001	1.73
Standard Commission - Own Given	4.10	0.32	1035.60	12.92	<.001	2.18
Standard Commission - Own Omission	4.05	0.47	1037.03	8.67	<.001	2.15
zOwn Correct - Standard Correct	-1.68	0.15	1030.83	-11.57	<.001	-0.89
Own Correct - Standard Given	-0.16	0.20	1031.23	-0.77	.994	-0.08
Own Correct - Standard Omission	-1.85	0.36	1039.91	-5.16	<.001	-0.99
Standard Correct - Own Given	2.52	0.26	1033.44	9.88	<.001	1.34
Standard Correct - Own Omission	2.47	0.43	1035.81	5.78	<.001	1.31
Own Given - Standard Given	-1.00	0.29	1032.05	-3.41	.016	-0.53
Own Given - Standard Omission	-2.69	0.42	1037.99	-6.48	<.001	-1.43
Standard Given - Own Omission	0.94	0.45	1034.71	2.09	.420	0.50
Own Omission - Standard Omission	-2.64	0.53	1032.11	-4.95	<.001	-1.40

Note. Standard refers to the correct diagram-standard. All bold rows were reported in the article.

Appendix J. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.cedpsych.2023.102251>.

References

- Anderson, M. C. M., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica*, 128(1), 110–118. <https://doi.org/10.1016/j.actpsy.2007.10.006>
- Baars, M., Vink, S., van Gog, T., de Bruin, A., & Paas, F. (2014). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learning and Instruction*, 33, 92–107. <https://doi.org/10.1016/j.learninstruc.2014.04.004>
- Bol, L., Riggs, R., Hacker, D., Dickerson, D., & Nunnery, J. (2010). The calibration accuracy of middle school students in match classes. *Journal of Research in Education*, 21, 81–96.
- Brand-Gruwel, S., Kammerer, Y., Van Meeuwen, L., & Van Gog, T. (2017). Source evaluation of domain experts and novices during Web search. *Journal of Computer Assisted Learning*, 33(3), 234–251. <https://doi.org/10.1111/jcal.12162>
- Dalpia, David. Applied Statistics with R, *GitHub* (2022 - 2023). <https://book.stat420.org/>.
- Carroll, L. (2015). *Alice's adventures in wonderland and through the looking glass*. Princeton University Press.
- de Bruin, A. B. H., 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>
- Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgements made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction*, 24, 4–14. <https://doi.org/10.1016/j.learninstruc.2012.06.001>
- Dunlosky, J., Hartwig, M. K., Rawson, K. A., & Lipko, A. R. (2011). Improving college students' evaluation of text learning using idea-unit standards. *Quarterly Journal of Experimental Psychology*, 64(3), 467–484. <https://doi.org/10.1080/17470218.2010.502239>
- Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, 22(4), 271–280. <https://doi.org/10.1016/j.learninstruc.2011.08.003>
- Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction*, 24, 58–61. <https://doi.org/10.1016/j.learninstruc.2012.05.002>
- Goodman, L.A., Kruskal, W.H. (1979). Measures of Association for Cross Classifications. *Springer Series in Statistics*. Springer, New York, NY. <https://doi.org/10.1007/978-1-4612-9995-0.1>
- Griffin, T. D., Mielicki, M. K., & Wiley, J. (2019). Improving students' metacomprehension accuracy. *The Cambridge Handbook of Cognition and Education*, 2, 1535299, 619–646. <https://doi.org/10.1017/9781108235631.025>
- Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition*, 36(1), 93–103. <https://doi.org/10.3758/mc.36.1.93>
- Hacker, D. J., & Bol, L. (2019). Calibration and Self-Regulated Learning. *The Cambridge Handbook of Cognition and Education*, 647–677. <https://doi.org/10.1017/9781108235631.026>
- Händel, M., & Dresel, M. (2018). Confidence in performance judgment accuracy: The unskilled and unaware effect revisited. *Metacognition and Learning*, 13(3), 265–285. <https://doi.org/10.1007/s11409-018-9185-6>
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.
- Jaeger, A. J., & Wiley, J. (2014). Do illustrations help or harm metacomprehension accuracy? *Learning and Instruction*, 34, 58–73. <https://doi.org/10.1016/j.learninstruc.2014.08.002>
- Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language*, 29(2), 133–159. [https://doi.org/10.1016/0749-596x\(90\)90069-c](https://doi.org/10.1016/0749-596x(90)90069-c)
- Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, 100(4), 609–639. <https://doi.org/10.1037/0033-295x.100.4.609>
- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgements of learning. *Journal of Experimental Psychology: General*, 126(4), 349–370. <https://doi.org/10.1037/0096-3445.126.4.349>
- Lachner, A., Burkhart, C., & Nückles, M. (2017). Mind the gap! Automated concept map feedback supports students in writing cohesive explanations. *Journal of Experimental Psychology: Applied*, 23(1), 29–46. <https://doi.org/10.1037/xap0000111>
- Lipko, A. R., & Dunlosky, J. (2007). Using Feedback to Improve Grade School Students Judgments of Text Learning. *PsycEXTRA Dataset*. <https://doi.org/10.1037/e658672007-001>
- Lipko, A. R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, K., & Cook, D. (2009). Using standards to improve middle school students' accuracy at evaluating the quality of their recall. *Journal of Experimental Psychology: Applied*, 15(4), 307–318. <https://doi.org/10.1037/a0017599>
- Alexander, P. A. (2015). Examining response confidence in multiple text tasks. *Metacognition and Learning*, 10, 407–436. <https://doi.org/10.1007/s11409-015-9138-2>
- Maki, R. H. (1998). Metacomprehension of Text. *Psychology of Learning and Motivation*, 223–248. [https://doi.org/10.1016/s0079-7421\(08\)60188-7](https://doi.org/10.1016/s0079-7421(08)60188-7)
- Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(4), 663–679. <https://doi.org/10.1037/0278-7393.10.4.663>
- McCrudden, M. T., Schraw, G., Lehman, S., & Poliquin, A. (2007). The effect of causal diagrams on text learning. *Contemporary Educational Psychology*, 32(3), 367–388. <https://doi.org/10.1016/j.cedpsych.2005.11.002>
- Moerbeek, M., & Teerenstra, S. (2016). *Power analysis of trials within multilevel data*. Boca Raton: CRC Press.
- Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the “labor-in-vain effect”. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(4), 676–686. <https://doi.org/10.1037/0278-7393.14.4.676>
- Prinz, A., Golke, S., & Wittwer, J. (2020a). How accurately can learners discriminate their comprehension of texts? A comprehensive meta-analysis on relative metacomprehension accuracy and influencing factors. *Educational Research Review*, 31(August). <https://doi.org/10.1016/j.edurev.2020.100358>
- Prinz, A., Golke, S., & Wittwer, J. (2020b). To what extent do situation-model-approach interventions improve relative metacomprehension accuracy? Meta-analytic insights. *Educational Psychology Review*, 32(4), 917–949. <https://doi.org/10.1007/s10648-020-09558-6>
- R Core Team. (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology*, 19(4–5), 559–579. <https://doi.org/10.1080/09541440701326022>
- 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>
- Posit team (2022). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. <https://support.posit.co/hc/en-us/articles/206212048-Citing-RStudio>.
- Strijbos, J. W., Martens, R. L., Prins, F. J., & Jochems, W. M. (2006). Content analysis: What are they talking about? *Computers & Education*, 46(1), 29–48. <https://doi.org/10.1016/j.compedu.2005.04.002>
- Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology*, 28(2), 129–160. [https://doi.org/10.1016/s0361-476x\(02\)00011-5](https://doi.org/10.1016/s0361-476x(02)00011-5)
- Thiede, K. W., Anderson, M. C. M., & Theriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, 95(1), 66–73. <https://doi.org/10.1037/0022-0663.95.1.66>
- Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(4), 1024–1037. <https://doi.org/10.1037/0278-7393.25.4.1024>
- Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. M. (2010). Poor Metacomprehension Accuracy as a Result of Inappropriate Cue Use. *Discourse Processes*, 47(4), 331–362. <https://doi.org/10.1080/01638530902959927>
- Thiede, K. L., Wright, K., Hagenah, S., & Wenner, J. (2019). Drawings as Diagnostic Cues for Metacomprehension Judgment. *Metacognition in Learning*. <https://doi.org/10.5772/intechopen.86959>.
- van de Pol, J., de Bruin, A. B. H., van Loon, M. H., & van Gog, T. (2019). Students' and teachers' monitoring and regulation of students' text comprehension: Effects of comprehension cue availability. *Contemporary Educational Psychology*, 56, 236–249. <https://doi.org/10.1016/j.cedpsych.2019.02.001>
- van de Pol, J., van Gog, T., & Thiede, K. (2021). The relationship between teachers' cue-utilization and their monitoring accuracy of students' text comprehension. *Teaching and Teacher Education*, 107, Article 103482. <https://doi.org/10.1016/j.tate.2021.103482>
- van de Pol, J., van Loon, M., van Gog, T., Braumann, S., & de Bruin, A. (2020). Mapping and Drawing to Improve Students' and Teachers' Monitoring and Regulation of Students' Learning from Text: Current Findings and Future Directions. *Educational Psychology Review*, 32(4), 951–977. <https://doi.org/10.1007/s10648-020-09560-y>
- van Gog, T. (2006). Uncovering the problem-solving process to design effective worked examples. [Doctoral dissertation, Open University of The Netherlands, Heerlen, The Netherlands]. https://research.ou.nl/ws/files/934720/Dissertation_summaries%20Van%20Gog%202006.pdf.
- van Gog, T., Paas, F., van Merriënboer, J. J. G., & Witte, P. (2005). Uncovering the Problem-Solving Process: Cued Retrospective Reporting Versus Concurrent and Retrospective Reporting. *Journal of Experimental Psychology: Applied*, 11(4), 237–244. <https://doi.org/10.1037/1076-898x.11.4.237>
- van Gog, T., & Scheiter, K. (2010). Eye tracking as a tool to study and enhance multimedia learning. *Learning and Instruction*, 20(2), 95–99. <https://doi.org/10.1016/j.learninstruc.2009.02.009>
- van Loon, M. H., de Bruin, A. B. H., van Gog, T., van Merriënboer, J. J. G., & 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>
- Waldeyer, J., & Roelle, J. (2020). The keyword effect: A conceptual replication, effects on bias, and an optimization. *Metacognition and Learning*, 16(1), 37–56. <https://doi.org/10.1007/s11409-020-09235-7>
- Winne, P. H., & Hadwin, A. F. (2010). Self-Regulated Learning and Socio-Cognitive Theory. *International Encyclopedia of Education*, 503–508. <https://doi.org/10.1016/b978-0-08-044894-7.00470-x>
- Zamary, A., Rawson, K. A., & Dunlosky, J. (2016). How accurately can students evaluate the quality of self-generated examples of declarative concepts? Not well, and feedback does not help. *Learning and Instruction*, 46, 12–20. <https://doi.org/10.1016/j.learninstruc.2016.08.002>