

This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Pijeira-Díaz, Héctor J.; van de Pol, Janneke; Channa, Faisal; de Bruin, Anique

Title: Scaffolding self-regulated learning from causal-relations texts : Diagramming and self-assessment to improve metacomprehension accuracy?

Year: 2023

Version: Published version

Copyright: © The Author(s) 2023

Rights: CC BY 4.0

Rights url: <https://creativecommons.org/licenses/by/4.0/>

Please cite the original version:

Pijeira-Díaz, H. J., van de Pol, J., Channa, F., & de Bruin, A. (2023). Scaffolding self-regulated learning from causal-relations texts : Diagramming and self-assessment to improve metacomprehension accuracy?. *Metacognition and Learning*, 18(3), 631-658.
<https://doi.org/10.1007/s11409-023-09343-0>



Scaffolding self-regulated learning from causal-relations texts: Diagramming and self-assessment to improve metacomprehension accuracy?

Héctor J. Pijeira-Díaz¹ · Janneke van de Pol² · Faisal Channa³ · Anique de Bruin¹

Received: 30 June 2022 / Accepted: 29 March 2023
© The Author(s) 2023

Abstract

The accuracy of students' relative comprehension judgments when reading texts is typically rather low. This has been ascribed to students grounding their comprehension judgments on cues that are not diagnostic of their actual comprehension level. Asking students to complete causal diagrams—a diagramming scaffold—before judging comprehension has proved effective in providing them with more diagnostic cues and thereby fostered metacomprehension accuracy and self-regulated learning. However, there is still room for improvement. We investigated experimentally whether adding the instruction to students to self-assess their causal diagrams: (1) would lead to more accurate judgments than comprehension judgments, (2) would boost their utilization of diagnostic diagram cues by increasing the saliency of those cues, and (3) would enhance metacomprehension accuracy. Participants ($N=427$ secondary students in The Netherlands) were randomly assigned to one of three conditions, namely (1) only diagram completion, (2) diagram completion plus diagram self-assessment, or a (3) filler task after reading (control). Self-assessments were more accurate than comprehension judgments, while both correlated strongly. However, no significant differences were found between diagramming conditions concerning diagram cue utilization and metacomprehension accuracy. Apparently, students self-assess their diagrams even without instruction to do so. Nonetheless, the effect of the diagramming scaffold for improving relative metacomprehension accuracy was replicated and extended to absolute metacomprehension accuracy.

Keywords Cue utilization · Diagramming · Metacomprehension accuracy · Scaffolding · Self-assessment · Self-regulated learning

✉ Héctor J. Pijeira-Díaz
h.pijeiradiaz@maastrichtuniversity.nl

¹ Department of Educational Development and Research, School of Health Professions Education (SHE), Faculty of Health, Medicine and Life Sciences (FHML), Maastricht University, Maastricht, The Netherlands

² Department of Education, Utrecht University, Utrecht, The Netherlands

³ Department of Teacher Education, University of Jyväskylä, Jyväskylä, Finland

Scaffolding self-regulated learning from causal-relations texts: Diagramming and self-assessment to improve metacomprehension accuracy?

Learning from texts is an important part of almost all school subjects. In the process of self-regulated learning from reading, students make comprehension judgments (Dunlosky & Lipko, 2007) and (re)study and strategy decisions (Thiede et al., 2009). According to theories of self-regulated learning, (in)accurate comprehension judgments are conducive to (in)effective (re)study decisions and suboptimal strategy choices (Butler & Winne, 1995; Dinsmore & Parkinson, 2013). Research has shown that students have great difficulties with accurately judging their understanding (Stone, 2000; van de Pol et al., 2019). The problem becomes then that erroneous comprehension judgments may lead to inefficient or insufficient allocation of effort and ultimately to deficient performance (Kanfer & Ackerman, 1989). Given the central role of learning from texts in education, understanding comprehension judgments has attracted substantial research (Dunlosky & Thiede, 2013; Händel et al., 2020).

Already a few decades ago, attention was drawn to the problem of inaccurate comprehension judgments, calling it “the illusion of knowing” or “failure in the self-assessment of comprehension” (Glenberg et al., 1982). Since then, tens of studies have consistently shown how inaccurate students’ judgments of their own understanding are, with a correlation between students’ judgments and their actual understanding of 0.27 reported as average across over 20 datasets (cp. Dunlosky & Lipko 2007; Maki, 1998), a figure deemed as “alarming” (Graesser, 2007, p. 4). After all, on the way to the goal of text understanding, why would one expect a learning regulation decision (e.g., use of strategies for comprehension or allocation of processing resources) to be judicious if it is based on faulty monitoring?

The pervasiveness of inaccurate comprehension judgments has been explained by research showing that students tend to base their judgments on cues which are not diagnostic or representative of actual comprehension (de Bruin et al., 2017; Koriat, 1997; Thiede et al., 2010). Consequently, generative interventions have been developed and evaluated to scaffold students in the accuracy of their comprehension judgments (hereinafter *metacomprehension accuracy*) by making diagnostic cues more salient and thereby available to them. In the last decade, a diagramming scaffold has shown effectiveness to enhance students’ metacomprehension accuracy (for a review, see van de Pol et al., 2020). In completing diagrams, students are faced with cues that are diagnostic of their comprehension (e.g., number of completed diagram boxes; see Fig. 1 for an example diagram with four empty boxes to be completed by the students). However, there is still room for improvement concerning students’ metacomprehension accuracy and utilization of the diagram cues (Prinz et al., 2020).

In this study, building on the diagramming paradigm (van de Pol et al., 2019; van Loon et al., 2014), we test whether asking students to self-assess their own diagrams (i.e., another scaffold) by indicating how many diagram boxes they judge to have completed correctly (see Fig. 1 for an example diagram), increases diagnostic cue utilization and thereby the accuracy of comprehension judgments, provided that self-assessments are accurate enough. The expected role of self-assessment scaffold is to focus students’ attention on the most diagnostic—yet underutilized—cue of number of correctly completed diagram boxes, which could elicit a deeper reflection on their own text understanding as a means for their self-assessments. In addition, self-assessments have been long and frequently used in the

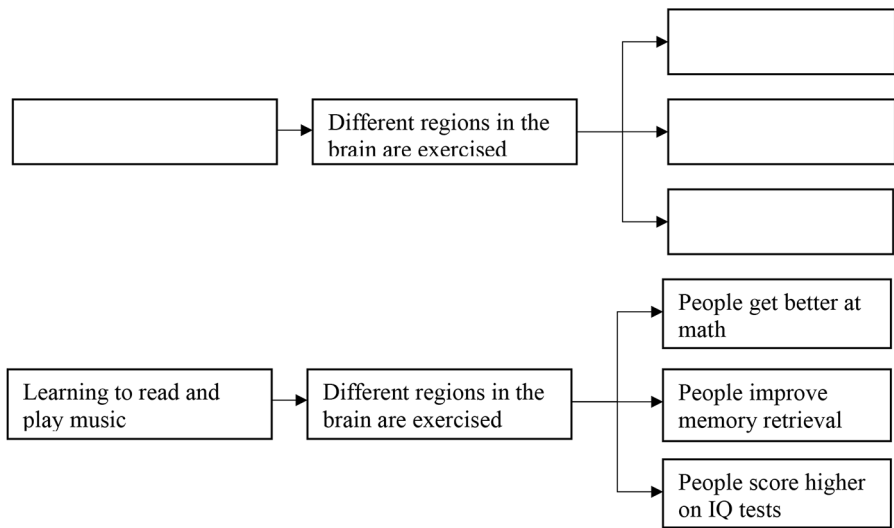


Fig. 1 Pre-structured diagram as presented to students (top) and correctly completed (bottom). Based on study materials from Van Loon et al. (2014).

classroom for a variety of purposes (Andrade, 2019; Gale, 1984) and are a scaffold that students can take on their own without need for a particular material or additional support from their teachers.

Next, we will first explain how cue diagnosticity and cue utilization affect metacomprehension accuracy, before turning to the diagramming scaffold and subsequently to previous results on students' self-assessments.

Cue utilization and cue diagnosticity

In their influential review, Butler and Winne (1995) argued that one of the main reasons why students have difficulty in monitoring their own comprehension pertains to students' cue-utilization. That is, students might use cues that lack validity for predicting performance (i.e., cue diagnosticity). The use of cues is central to the predominant view that metacognitive judgments are inferential in nature (Bjork et al., 2013; Dunlosky et al., 2020), as opposed to early assumptions (Hart, 1965) that people have direct access to a sort of internal monitor, from where information on memory content is readily available (Koriat, 2012).

Both (perceived) task characteristics and students' progress on tasks are described as a set of features or a profile of cues (Butler & Winne, 1995). Students weight the cues (e.g., domain knowledge, mental effort, and/or remaining time) to predict their performance in an analogous way to a multiple regression model, thereby simultaneously integrating multiple cues (Brunswik, 1956). Cue utilization for a metacognitive judgment is then determined for each cue by the correlation between the cue and the judgment (Koriat, 2012).

Cues, however, differ in their diagnosticity or extent to which they predict performance. Historically, this notion has been expressed via different terminology such as "ecological validity" or "trustworthiness" of a cue (Brunswik, 1956), "cue relevance" (Nystedt & Magnusson, 1973), "cue reliability" (York et al., 1987), and "cue validity" (Balzer et al., 1989).

Metacomprehension accuracy varies accordingly with the diagnosticity of the cues used for metacognitive judgments (Prinz et al., 2020). Therefore, cue utilization should be contingent on cue diagnosticity if judgment accuracy is to be maximized (Brunswik, 1956; van de Pol et al., 2020). That is, people should use diagnostic cues to produce accurate judgments. However, when monitoring text comprehension, students generally use cues of low diagnosticity such as domain familiarity and interest in the topic (Thiede et al., 2010). Empirical studies show, on the one hand, that producing content related to a certain text (e.g., summaries and diagrams), generates cues of the highest diagnosticity (i.e., predictive power) of performance on comprehension tests; but on the other hand, that, unfortunately, students rarely use those cues (Thiede et al., 2010).

The diagramming scaffold for metacomprehension

A direct application of cue-utilization theory and the available evidence on diagnostic cues is that scaffolds can be designed to guide students towards generating and/or using more reliable cues that enhance their metacognitive judgments and thereby their self-regulated learning (Björkman, 1994; Dunlosky & Thiede, 2013; van de Pol et al., 2020). Tested scaffolds in the form of generative tasks known to enhance monitoring of reading comprehension have included listing keywords (de Bruin et al., 2011; Thiede et al., 2003), summarizing (Thiede et al., 2010; Thiede & Anderson, 2003), concept mapping (Redford et al., 2012; Thiede et al., 2010) and more recently, completing pre-structured diagrams (van de Pol et al., 2019; van Loon et al., 2014).

In developing the diagramming paradigm (i.e., read texts+use diagramming scaffold+make comprehension judgments+choose texts to restudy+take a comprehension test; see Fig. 2), Van Loon et al. (2014) adapted the metacomprehension paradigm by introducing a diagnostic-cue-generation scaffold in the form of a pre-structured diagram (see Fig. 1 for an example) between text reading and comprehension judgments. After making the comprehension judgments, students are prompted with the metacognitive control or self-regulation decision of which text(s) they would like to restudy, if any.

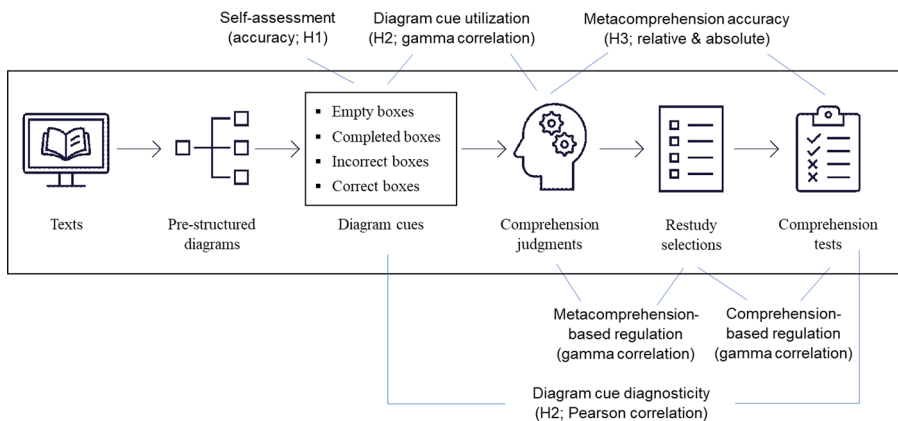


Fig. 2 The diagramming scaffold for metacomprehension accuracy with the constructs involved in hypotheses H1, H2, and H3 as well as the others

Scaffolding students with pre-structured diagrams to be completed is particularly well-suited to expository texts (as opposed to narrative texts). Expository texts are commonplace for example in science, geography, and history subjects where the storyline is supported by causal relations (Fiorella & Mayer, 2016; Wiley et al., 2005). In general, expository texts are harder to comprehend (Williams, 2007), which by extension might complicate metacomprehension accuracy. Hence, a focus and context for the applicability of the pre-structured diagram completion scaffold.

To further improve and understand better its working mechanism, the diagramming scaffold has been tested experimentally in terms of timing (i.e., immediately after reading a text vs. after a delay) (van Loon et al., 2014) and level of support (i.e., completing a pre-structured diagram vs. drawing the diagram from scratch) (van de Pol et al., 2019). Aligned with the delayed-judgment effect (Rhodes & Tauber, 2011), it was found that students who completed the diagrams after a delay outperformed those who did it immediately after reading each text in terms of both metacomprehension accuracy and metacomprehension-based regulation (van Loon et al., 2014; see Fig. 2 for the constructs). Regarding level of support, however, no statistically significant difference was found between either completing or drawing the diagrams and either metacomprehension accuracy or regulation (van de Pol et al., 2019).

Probably because students do not actually restudy the texts that they select to, no statistically significant differences have been found in test scores between students in the diagramming and no-diagramming conditions (see Table 1). Also contained in Table 1 are the results regarding both cue diagnosticity (i.e., traditionally determined by the Pearson correlation between each diagram cue and the test scores) and cue utilization (i.e., tradition-

Table 1 Diagramming paradigm findings (statistical means) for causal relations in expository texts (no diagram and delayed diagram conditions)

	Van Loon et al. (2014)	Van de Pol et al. (2019)
No diagram → Delayed diagram (<i>p</i> contrast)		
Relative metacomprehension accuracy (−1.0 to 1.0; gamma correlation)	0.07 → 0.56 (<i>p</i> = .005)	0.00 → 0.43 (<i>p</i> = .002)
Metacomprehension-based regulation (−1.0 to 1.0; gamma correlation)	−0.69 → −0.74 (<i>p</i> = .984)	−0.68 → −0.76 (<i>p</i> = .700)
Comprehension-based regulation (−1.0 to 1.0; gamma correlation)	N/A	−0.17 → −0.31 (<i>p</i> = .114)
Test score (relations) (0–4)	1.77 → 1.97 (<i>p</i> = .175)	1.20 → 1.17 (<i>p</i> = .163)
Cue diagnosticity (Pearson correlation)		
Omission errors	−0.39	−0.26
Completed boxes	N/A	0.26
Commission errors	−0.17	−0.14
Correct boxes	0.49	0.39
Cue utilization (Gamma correlation)		
Omission errors	−0.64	−0.71
Completed boxes	N/A	0.71
Commission errors	−0.16	0.10
Correct boxes	0.59	0.57

N/A: not available

ally calculated as the gamma correlation between each diagram cue and the comprehension judgments; see these constructs in context in Fig. 2) for four diagram cues, namely, number of omission errors (i.e., diagram boxes left empty), number of completed boxes (i.e., the opposite of omissions), number of commission errors (i.e., diagram boxes completed incorrectly), and the number of correct boxes (i.e., diagram boxes completed correctly). As can be observed, “number of correct boxes” is the cue of highest diagnosticity, while “number of omissions errors” (together with its opposite, number of completed boxes) seems to be more utilized. Note that cue utilization is calculated using correlations, so causality cannot be asserted.

Diagramming increases the availability of diagnostic cues (van de Pol et al., 2020). However, despite the clear beneficial effect of the diagramming scaffold on metacomprehension accuracy (see Table 1), further improvement is warranted. Promoting students’ utilization of diagnostic cues might be a key to such enhancement. In this work, we investigate whether students’ self-assessments of their own diagrams contribute to narrowing the gap between cue diagnosticity and cue utilization. With such self-assessments students are asked, as a scaffold, to indicate how many diagram boxes they have (1) completed and (2) completed correctly.

Students’ self-assessment of their answers

Students’ self-assessments have a long tradition in educational psychology, either in the classroom across educational levels, for distance learning, or for independent learning (Andrade, 2019; Boud & Falchikov, 1989). Different terms used for self-assessments are “self-scoring”, “self-marking”, “self-rating”, “self-evaluation”, and “self-diagnosis”.

According to Gale’s (1984) taxonomy, an important purpose of using self-assessment scaffolds entails the diagnosis and self-remediation of learning problems, which in the light of self-regulated learning theory could be interpreted as the reciprocal interaction between monitoring and control (Nelson & Narens, 1990). In her review, Gale (1984) includes other applications of self-assessment of which the most related to this study are to emphasize key issues, to increase students’ attention to the object of self-assessment (e.g., in our case, the number of correct diagram boxes as a diagnostic cue for students’ comprehension judgments), and as retrospective organizers of prior learning.

In this study, we use the self-assessment scaffold as a means to focus learners specifically on the most diagnostic cue(s) of their diagrams, by making them more salient, as a step towards increasing the use of these cues when judging their comprehension. This is expected to improve their metacomprehension accuracy. More specifically, according to Koriat et al. (1980), the mechanism for making judgments of learning (JOLs, or comprehension judgments in their projection to our case), consists of three stages, namely, (1) searching memory for relevant evidence, (2) judging the evidence, and (3) translating the evidence into a number. We consider Brunswik’s (1956) notion of the integration of multiple cues in a fashion that is analogous to a multiple regression model to be part of the first stage. If we ask students to self-assess their diagrams and make a comprehension judgment right thereafter, based on this line of reasoning, we are influencing stages one and two, since both the completed diagram and its self-assessment are fresh in students’ memory and are thereby available in stage three of expressing the comprehension judgment as a more accurate number.

Previous research provides evidence that students are already able to use the diagnostic cues of omissions and number of correct boxes to a certain extent when judging their comprehension. In the study by Van Loon et al. (2014), the utilization of the omissions cue was -0.64 and that of correct boxes was 0.59 (see Table 1). Our assumption is that by having students explicitly self-assess these diagnostic cues, they become more salient, the use of these cues increases and hence metacomprehension accuracy improves. Even though their self-assessments will not always be accurate, based on previous research it is reasonable to assume that students should be able to profit to a certain extent from making explicit self-assessments.

The accuracy of self-assessments has long been of interest as well (Gale, 1984). Mixed results are found in reviews of self-assessment studies (Andrade, 2019; Brown & Harris, 2014; Yan et al., 2022), but the type and purpose of self-assessment scaffolds should be used as determining criteria for comparable findings. In the initial metacomprehension paradigm, Glenberg and Epstein (1985) observed that, after reading a text and answering an inference question, subjects were accurate in judging the correctness of their answers. Thereafter, students' self-assessment of their answers has been consistently found more accurate than their comprehension judgments (Glenberg & Epstein, 1987; Huff & Nietfeld, 2009; Maki et al., 1994; Ramdass & Zimmerman, 2008; Walczyk & Hall, 1989). Such stable effect has been explained by the vantage point of being able to rely on the experience with the task, since the assessment occurs a posteriori (i.e., after taking the test or, in this case, completing the diagram) (Labuhn et al., 2010).

In sum, we expect the self-assessment scaffold to contribute to enhancing metacomprehension accuracy by increasing the saliency of diagnostic cues when making confidence judgments (Glenberg et al., 1987; Huff & Nietfeld, 2009).

Present study

This work built on the diagramming paradigm (van Loon et al., 2014) where participants (1) read six texts containing a number of facts and causal relations, (2) complete diagrams based on the texts' causal relations (or a filler task in the control condition), (3) judge their comprehension for each text's facts and causal relations (i.e., comprehension judgments), (4) select texts for restudy (without actually restudying them), and (5) are tested on each text for their reading comprehension (comprising both the facts and causal relations in the texts).

Two different types of metacomprehension accuracy have been explored in metacomprehension research, namely, relative (also called "resolution" or "discrimination") and absolute (also called "calibration" or "bias"; Dunlosky & Lipko 2007). Since the two kinds of accuracy are statistically independent, we follow the recommendation in the field to report both (Dunlosky & Thiede, 2013). In line with reporting recommendations for absolute metacomprehension accuracy (Nelson & Dunlosky, 1991; Pieschl, 2009; Stone, 2000), the calibration curves for facts and causal relations are computed.

The diagramming paradigm allows for estimating the effect of metacognitive monitoring on metacognitive control in that context, which we refer to as "metacomprehension-based regulation", and also to compute the "comprehension-based regulation" (i.e., association between restudy selections and actual test performance; cf. van de Pol et al., 2020). The calculation of regulation measures answers the call to account for metacognitive control

processes when studying metacognitive judgments (Pieschl, 2009), since monitoring and control processes inform each other.

In the diagram completion scaffold, the number of correct boxes has been found to be the most diagnostic cue but also not the most utilized cue by the students. Obviously, students do not have access to the correctness of their completed diagram. The closer they can get is by self-assessing their diagrams. By adding a self-assessment scaffold in the form of prompts to the diagram completion scaffold, we expect to draw students' attention towards the correct boxes (the most diagnostic cue).

We contrast the effects of a diagram completion (DC), a diagram completion plus self-assessment (DC+SA), and a filler-task control group (C) on students' metacomprehension accuracy regarding the causal relations in the texts. Figure 2 shows the constructs involved in the hypotheses in the context of the diagramming paradigm. Our three hypotheses are as follows.

H1 Self-assessment hypothesis: Self-assessments are more accurate than comprehension judgments and they predict comprehension judgments to some extent in the diagramming paradigm.

In a variety of related literature, self-assessments have been repeatedly shown to be more accurate than comprehension judgments, which is sometimes referred to as the "postdiction superiority effect" (Pierce & Smith, 2001). We expect these results to extend to the diagramming paradigm. This hypothesis also serves as an assumption for the following ones, as we intend to use a presumably more accurate judgment (i.e., self-assessment) to improve a presumably less accurate one (i.e., comprehension judgment). In addition, we would like to find out the extent to which they are correlated, as a strong relationship could mean that influencing one judgment would have a reflection on the other, thereby providing more opportunities for enhancing judgment accuracy. Similarly, we are interested in whether the accuracy of self-assessments predicts the accuracy of comprehension judgments.

H2 Diagnostic-cue-utilization hypothesis: Students who self-assess their diagrams utilize the diagnostic cues generated at diagram completion more when monitoring text learning compared to students who only complete the diagram.

This is, diagnostic cue utilization for $DC+SA > DC$. This hypothesis builds on the previous finding that the number of correctly completed diagram boxes is the most diagnostic cue generated with the diagramming scaffold (van de Pol et al., 2020). Through asking students to self-assess their correctly completed diagram boxes, we expect to focus students' attention on this most diagnostic cue and thereby that they increase its utilization when making their comprehension judgments.

H3 Metacomprehension-accuracy hypothesis: Students' metacomprehension accuracy (both relative and absolute) will be higher when they self-assess their diagrams than only diagramming, and in turn, higher for diagramming only than no diagramming (i.e., control).

This is, metacomprehension accuracy for $DC+SA > DC > C$. It has been experimentally shown (e.g., van de Pol et al., 2019; van Loon et al., 2014) that DC significantly enhances relative metacomprehension accuracy over no diagramming (i.e., C). We

expect that DC+SA will further increase metacomprehension accuracy by drawing students' attention to diagnostic but underutilized cues as indicated by the previous studies when analyzing accuracy shortcomings. Compared to relative accuracy, absolute metacomprehension accuracy has been significantly less investigated (Prinz et al., 2020). Nonetheless, we expect that in both conditions, DC+SA and DC, the relative accuracy improvement extends also to absolute metacomprehension accuracy, given the more direct connection that the latter represents between judgment and performance (i.e., operationalized by the simple absolute deviation). In a review by Van de Pol et al. (2020), four of the five drawing (i.e., concepts maps or other schemas) studies analyzing absolute metacomprehension accuracy reported that the drawing intervention enhanced absolute accuracy.

For completeness and comparability with previous studies, we contrast the previous aspects regarding factual information in addition to the causal relations in the text. However, since the diagram scaffold targets causal relations and not facts, we do not expect that the previous hypotheses hold for facts. Similarly, although outside our hypotheses for this study, we supplementarily analyze other important indicators customarily reported in the diagram paradigm such as those related to the regulation (or control) that students exercise when making restudy selections and to reading comprehension as such, as determined by the comprehension test. In this way, we provide a holistic view of the diagramming paradigm indicators capturing different facets of the learning process.

Method

Three experimental conditions were employed in a between-subjects design that consisted of five steps as depicted in Fig. 3.

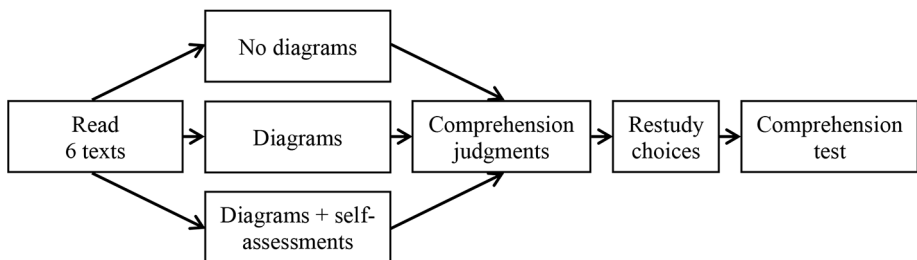


Fig. 3 Experimental design

Participants

Participants ($N=427^1$) were secondary education students between 11 and 17 years of age ($M=14.54$, $SD=1.06$). The gender distribution was 213 females, 204 males, and 10 unknown (students who did not indicate their gender in the questionnaire). Most learners were Dutch nationals ($n=415$), while 12 were of other nationalities. Students were not asked to indicate their mother tongue, but they were all following their education in regular Dutch classrooms. Therefore, their level of Dutch was deemed as sufficient for the experiments. Participation in the study was voluntary. Students gave informed written consent while parents gave passive consent. In other words, parents could object if they did not agree that their child would participate in the study. We complied with the APA ethical standards for treatment of human participants, informed consent, and data management.

Within school classes, participants were randomly divided into three groups corresponding to the experimental conditions: (1) DC ($n=145$), (2) DC+SA ($n=143$), and (3) C ($n=139$). There were no statistically significant differences (at the 95% confidence interval) among groups in terms of gender ($p=.863$), educational level ($p=.978$), nationality ($p=.721$), or age ($p=.440$). Conversely, there were significantly ($p<.001$) more students with dyslexia in the control group ($n=41$) than in the diagram completion ($n=27$) and in the diagram completion plus self-assessment ($n=16$) groups.

Materials

Seven booklet types were used for the experiment, one for each experimental step (adding up to five, namely, one with the six texts to read, one for the experimental or control treatment, one for the comprehension judgments, one for the restudy decisions, and one with the comprehension test; see Fig. 3), an initial one for practice and a final one with puzzles as pastime to prevent students from leaving their seats until all of them had finished the experimental booklets.

Since there were six texts, six versions of each experimental booklet type were made to counterbalance the order of the topics, using a Latin Square design (e.g., one version would have the order Text1-Text2-Text3-Text4-Text5-Text6, the next one Text2-Text3-Text4-Text5-Text6-Text1, and so on). Texts and diagrams were placed one to a page and with a blank page between two of them.

Practice materials

The practice booklet contained general instructions and descriptions of the experiment (e.g., “work at your own pace”), two example texts, example questions about factual information in the texts, example questions about causal relations in the text, and an example of the format used to give comprehension judgments.

¹ 444 students participated in the experiment, but due to data issues, the researchers decided to exclude the data from 17 of them. Examples of such issues were missing experimental booklets and the reporting of invalid values maybe because those students did not understand (parts of) the task properly or did not take the task seriously (e.g., indicating self-assessment values of 48, 96, or -3, when the possible range was 0–4).

Texts

We used the same six expository causal-relations texts previously used in studies of the diagramming scaffold (e.g., van de Pol et al., 2019; van Loon et al., 2014). The topics are “Sunken Subway Cars”, “The Use of Botox”, “The Suez Canal”, “The Benefits of Music”, “Money and Happiness”, and “Renovation of Concrete Buildings”. The texts are written in Dutch and range in length from 158 to 186 words ($M = 169.3$; $SD = 10.7$). Each text contains four causal relations, which are the focus of the scaffold, and five facts. As examples, the English translation of two of the texts is available in [Appendix A](#). The initial practice texts were about “Sport is Healthy” (95 words) and “Suburbs” (111 words).

Experimental tasks

In the DC group, pre-structured diagrams were provided with five diagram boxes (four empty and one pre-filled; see Fig. 1 for an example). The experimental task in the DC+SA group was similar but with an extra page in the booklet containing the self-assessment scaffold, namely, 12 (2 SA questions/diagram x 6 diagrams) self-assessment questions. The diagram self-assessment scaffold read: “Now look back at the diagrams you filled in. For each diagram, enter how many boxes you have completed in total and how many boxes you think are correct. (Boxes are filled in correctly if you mentioned there the relationships from the text.) ATTENTION: you are not allowed to change the diagrams or write anything else on the booklet with the diagrams!”. Then, on the same page, diagram self-assessment was prompted for each text as follows: “Diagram [TEXT_NAME]. How many boxes did you complete? How many boxes did you complete correctly?” In the control group, a picture matching task was used as filler task. Each pair of images was related to the topics (e.g., a ship crossing the Suez Canal, a concrete wall, and a metro station). To ensure that participants received the same information in all three conditions, the same statement provided in the pre-filled diagram box in the diagramming groups was provided to the control group before the picture matching task (on the same page corresponding to each text title).

Comprehension judgments and restudy selections

Students made the comprehension judgments for each text separately for facts (0–5; spaced number list) and causal relations (0–4; spaced number list). Comprehension judgments were prompted as: “How many questions do you think you will answer correctly on the test about the text [TEXT_NAME]? On a subsequent page, the six text titles were listed next to a checkbox for students to mark which text(s) they wanted to restudy, if any. The question read: “Which text(s) would you like to read again before taking the test?”

Reading comprehension test

Finally, the test questions asked for the four causal relations and five facts in each text. An example of a test question for the causal relations in the “Suez Canal” text was: “The distance for ships navigating between Jeddah and Rotterdam has been significantly shortened. Why? Give your answer as complete as possible. Mention 4 causal relations.”

Procedure

The experiment was conducted in classrooms. Each student received an envelope with the separate booklets described above for each step on one corner of their table.

They started with a simultaneous practice exercise to get familiar with the procedure and the tasks. The practice exercise involved reading two texts. After reading the first practice text, students took a test on the facts and causal relations from the text. Then, a diagram for the causal relations in the text was drawn on the blackboard by research assistants as a demonstration. Once students read the second practice text, they were given 1 min to draw a diagram for the causal relations. After that, the correct diagram was drawn on the blackboard. To finalize the practice, students were shown (in the same practice booklet) how to indicate their comprehension judgments.

After the practice, students were asked to complete one booklet at a time at their own pace and to place the booklets when completed on the other corner of their tables. Once there, the research assistants picked up the booklets to prevent participants to look back at them. The experiment lasted about 1 h, in addition to 15 min of instructions and practice.

Coding students' diagrams and tests

Given the volume of data and the availability of research assistants to code students' answers, the data were processed in two waves, first the data for 210 students and then the data for the remaining 217. In each wave, a different pair of research assistants worked first on coding 20% of the data to ensure satisfactory interrater reliability, and then one of them coded the remaining 80%. In the first wave of data processing and coding, the interrater reliability surpassed Cohen's $\kappa=0.70$. In the second wave, the interrater reliability went over Cohen's $\kappa=0.83$. Therefore, the respective coders achieved sufficient interrater reliability in both waves (Cohen, 1988).

The diagram content was analyzed by first coding each diagram box as completed (i.e., some content was written) or as an omission (i.e., no response was provided). Completed boxes were in turn coded as correct (i.e., an element of the causal relation structure contained in the texts) or as a commission error (i.e., incorrect answer). Commission errors included information not contained in the texts or factual information. Repeated information such as rephrasing the content of the pre-filled box was also coded as commission error. Students' test answers were scored according to the number of correct facts (0–5) and the number of correct causal relations (0–4).

Operationalization of measures

Relative metacomprehension accuracy was operationalized by the intra-individual Goodman-Kruskal's (1954) gamma correlation² between comprehension judgments and actual test scores, following the vast majority of the metacomprehension literature (Dunlosky et al., 2011; van de Pol et al., 2020). The closer to +1, the higher the relative metacomprehension accuracy.

For *absolute metacomprehension accuracy*, we used the two most common indicators or operationalizations, namely, (1) bias and (2) deviation or accuracy (Pieschl, 2009). Both

² Gamma correlations are also referred to as discrimination or resolution accuracy (Nelson, 1984).

measures are based on the mean difference between comprehension judgments and the corresponding test scores, but they diverge in whether the difference is signed (bias) or unsigned (deviation). Accordingly, they measure different aspects of absolute accuracy and complement one another. Bias allows for determining over- and underconfidence. Positive values of bias indicate (the degree of) overconfidence (i.e., students' comprehension judgments are higher than their actual test results, and the extent thereof), and conversely, negative values indicate (the degree of) underconfidence. A calibration curve is customarily used to visualize bias (Stone, 2000). A calibration curve contains the average of actual test scores for each value on a comprehension judgment scale. A recommended supplementary aid to the calibration curve, which we therefore included, is the line of perfect calibration (i.e., 100% accuracy; Dunlosky & Thiede 2013; Weingardt et al., 1994).

Since with bias positive and negative values cancel each other out as it is a signed difference, deviation is a more proper measure for absolute metacomprehension accuracy. The smaller the deviation value (i.e., the closer to zero), the higher the absolute metacomprehension accuracy. In this paper, unless otherwise stated, we use absolute metacomprehension accuracy to refer to the deviation of the judgments with respect to the test scores.

Metacomprehension-based regulation was determined by the intra-individual gamma correlation between a student's comprehension judgments and their binary restudy selections (0=text not selected for restudy, 1=text selected for restudy), separately for facts and causal relations. The closer to -1, the more precise the metacomprehension-based regulation.

Comprehension-based regulation was estimated by the intra-individual gamma correlation between a student's binary restudy selections (0=text not selected for restudy, 1=text selected for restudy) and their test scores, separately for facts and causal relations. The closer to -1, the more precise the comprehension-based regulation.

Regarding self-assessments, measures were computed analogously as for absolute metacomprehension accuracy and aligned with previous operationalizations in the literature (e.g., Kostons et al., 2012), including bias and deviation (i.e., accuracy) for the self-assessment of completed and correct boxes. Bias is therefore the mean difference between the self-assessed number of completed or correct boxes and the number of actual completed or correct boxes, respectively. Similarly, as for absolute metacomprehension accuracy, positive bias values indicate overconfidence (i.e., students' self-assessments were higher than their actual diagramming results), and conversely, negative values indicate underconfidence. The calibration curve was also computed for self-assessments. In terms of deviation, the smaller the value (i.e., the closer to zero), the higher the self-assessment accuracy.

Following previous diagram paradigm studies (e.g., van de Pol et al., 2019; van Loon et al., 2014), *cue utilization* was determined for each student by the gamma correlation between each diagram cue (i.e., omissions, completed boxes, commission errors and correct boxes) and the corresponding comprehension judgment.

Also aligned with the previous diagram paradigm studies, *cue diagnosticity* was estimated for each student by the Pearson correlation between each diagram cue (e.g., number of commission errors) and the test scores. For both cue utilization and cue diagnosticity, the values should be interpreted depending on the particular cues as follow. For completed boxes and correct boxes, values closer to +1 indicate both higher utilization and higher diagnosticity of those cues. Conversely, for omissions and commission errors, values closer to -1 indicate both higher utilization and higher diagnosticity of those cues.

Table 2 Intraclass correlation coefficient (ICC) for the study variables at the judgment, student, and class levels

Variable	ICC Judgment	ICC Student	ICC Class
Relative metacomprehension accuracy: Facts	N/A	0.37	0.02
Relative metacomprehension accuracy: Relations	N/A	0.37	0.02
Absolute metacomprehension accuracy: Facts	0.97	0.30	0.02
Absolute metacomprehension accuracy: Relations	0.74	0.15	0.02
Metacomprehension-based regulation: Facts	N/A	0.29	0.02
Metacomprehension-based regulation: Relations	N/A	0.33	0.01
Comprehension-based regulation: Facts	N/A	0.46	0.02
Comprehension-based regulation: Relations	N/A	0.44	0.04

N/A: not available

Analyses

Given the hierarchy or nesting within the data (i.e., several judgments within students and several students from a class), we computed the intraclass correlation coefficient (ICC) to determine whether it was necessary to run multilevel analyses instead of ANCOVAs to test the hypotheses (Peugh, 2010). The ICCs are listed in Table 2. In general, $ICC > 0.05$ indicates an amount of clustering and non-independence requiring the use of multilevel modeling (Arend & Schäfer, 2019). For the variables in Table 2, the judgment level is only applicable for absolute metacomprehension accuracy since for the other variables the judgments are aggregated at the student level by the gamma correlations. For the variables where the judgment level is not applicable, we can see from Table 2 that $ICC < 0.05$ at the class level, leaving only the student level. In those cases, ANCOVA with dyslexia as covariate was used (see Appendix B for SPSS syntax) because dyslexia was the only demographic with significant differences across experimental groups (see *Participants*). Bonferroni adjustments were employed to control for multiple comparisons in the post-hoc tests. T-tests were used to check for differences in cue diagnosticity and utilization across the two diagram groups for each of the diagram cues. In addition, a t-test was used to compare the accuracy of self-assessments to that of comprehension judgments. The relationship between self-assessments and comprehension judgments was tested using Pearson correlation analysis (two-tailed).

When it comes to absolute metacomprehension accuracy, however, ICC was above 0.05 in at least two levels. Accordingly, we ran multilevel analyses considering two levels (i.e., judgment and student) for absolute metacomprehension accuracy for facts and causal relations. Multilevel regression was used for modelling how self-assessment accuracy of correct boxes predicts absolute metacomprehension accuracy for both facts and relations. All multilevel analyses were performed using Mplus (Muthén & Muthén, 2017) version 8.3.

Table 3 Means (and standard deviations) of the study variables per condition

	Control	Diagram completion	+ Self-assessment
Completed boxes (0–4)	N/A	2.86 (1.32)	2.99 (1.3)
Omission errors (0–4)	N/A	1.14 (1.32)	1.01 (1.3)
Commission errors (0–4)	N/A	1.28 (1.09)	1.32 (1.09)
Correct boxes (0–4)	N/A	1.57 (1.2)	1.67 (1.2)
Self-assessment: Completed boxes (0–4)	N/A	N/A	3.17 (1.26)
Self-assessment: Correct boxes (0–4)	N/A	N/A	2.28 (1.32)
Self-assessment bias: Completed boxes (-4–4)	N/A	N/A	0.15 (0.62)
Self-assessment bias: Correct boxes (-4–4)	N/A	N/A	0.58 (1.13)
Self-assessment accuracy: Completed boxes (0–4)	N/A	N/A	0.18 (0.62)
Self-assessment accuracy: Correct boxes (0–4)	N/A	N/A	0.85 (0.93)
Comprehension judgments: Facts (0–5)	2.74 (1.18)	2.41 (1.34)	2.52 (1.33)
Comprehension judgments: Relations (0–4)	2.08 (1.06)	2.03 (1.17)	1.98 (1.16)
Restudy selections (texts/student) (0–6)	2.15 (1.25)	2.61 (1.14)	2.71 (1.30)
Test score: Facts (0–5)	1.63 (1.13)	1.45 (1.11)	1.51 (1.11)
Test score: Relations (0–4)	1.42 (1.10)	1.47 (1.17)	1.50 (1.18)
Relative metacomprehension accuracy: Facts (-1.0 to 1.0)	0.11 (0.65)	0.15 (0.65)	0.15 (0.59)
Relative metacomprehension accuracy: Relations (-1.0 to 1.0)	0.11 (0.64)	0.36 (0.58)	0.32 (0.61)
Metacomprehension bias: Facts (-5 to 5)	1.12 (1.55)	0.96 (1.52)	1.02 (1.52)
Metacomprehension bias: Relations (-4 to 4)	0.66 (1.43)	0.56 (1.27)	0.49 (1.27)
Absolute metacomprehension accuracy: Facts (0–5)	1.53 (1.15)	1.39 (1.14)	1.45 (1.11)
Absolute metacomprehension accuracy: Relations (0–4)	1.24 (0.98)	1.03 (0.93)	0.99 (0.93)
Metacomprehension-based regulation: Facts (-1.0 to 1.0)	-0.6 (0.64)	-0.67 (0.55)	-0.78 (0.46)
Metacomprehension-based regulation: Relations (-1.0 to 1.0)	-0.54 (0.70)	-0.71 (0.52)	-0.70 (0.51)
Comprehension-based regulation: Facts (-1.0 to 1.0)	-0.15 (0.70)	-0.16 (0.67)	-0.18 (0.72)
Comprehension-based regulation: Relations (-1.0 to 1.0)	-0.20 (0.72)	-0.33 (0.67)	-0.32 (0.69)

N/A: not available

Results

Table 3 contains the means and standard deviations of the study variables for each of the experimental groups. When applicable, the information is presented separately for facts and causal relations.

Self-assessment accuracy (H1)

We hypothesized, based on the literature, that self-assessments would be more accurate than comprehension judgments and we wanted to find out the strength of the relationship between self-assessments and comprehension judgments. The students had a self-assess-

Table 4 Parameter estimates from the multilevel regression separate analyses for facts and relations with self-assessment accuracy of correct boxes as independent variable and absolute metacomprehension accuracy for facts and relations as dependent variables

Model	<i>B</i>	<i>SE B</i>	<i>p</i>
Facts	0.181	0.037	<0.001
Relations	0.239	0.041	<0.001

Table 5 Diagnosticity (Pearson correlation) and utilization (gamma correlation) of diagram cues with t-test contrast results between conditions

	Diagram completion		+ Self-assessment		t-test	
	Facts <i>M (SD)</i>	Relations <i>M (SD)</i>	Facts <i>M (SD)</i>	Relations <i>M (SD)</i>	<i>p</i> Facts	<i>p</i> Relations
Cue diagnosticity						
Completed boxes	0.06 ^a (0.47)	0.37 (0.41)	0.07 ^a (0.46)	0.34 (0.38)	0.894	0.557
Omission errors	-0.06 ^a (0.47)	-0.37 (0.41)	-0.07 ^a (0.46)	-0.34 (0.38)	0.894	0.557
Commission errors	-0.02 ^a (0.45)	-0.14 (0.47)	0.08 (0.43)	-0.21 (0.46)	0.063	0.176
Correct boxes	0.09 (0.48)	0.54 (0.37)	0.03 ^a (0.47)	0.54 (0.39)	0.235	0.998
Cue utilization						
Completed boxes	0.52 (0.57)	0.57 (0.54)	0.42 (0.66)	0.53 (0.56)	0.278	0.657
Omission errors	-0.52 (0.57)	-0.57 (0.54)	-0.42 (0.66)	-0.53 (0.56)	0.278	0.657
Commission errors	0.14 (0.58)	0.18 (0.62)	0.06 ^a (0.64)	-0.02 ^a (0.69)	0.340	0.014
Correct boxes	0.42 (0.54)	0.42 (0.56)	0.39 (0.58)	0.48 (0.52)	0.576	0.438

^a Correlation NOT significantly different from 0; $p > .103$.

ment accuracy for completed boxes of 96%³ and for correct boxes of 79%. Regarding absolute metacomprehension accuracy, the average values were 69%, 74% and 75% for the C, DC, and DC+SA conditions respectively. Therefore, as expected, the self-assessments of correct boxes were significantly more accurate than comprehension judgments for causal relations, $t(1605) = -2.993$, $p = .003$, Cohen's $d = -0.150$.

There was a strong and statistically significant correlation ($r = .64$) between the self-assessment of correct boxes and the comprehension judgments for causal relations. In other words, self-assessments explained 41% ($R^2 = 0.41$) of the variance in comprehension judgments.

Next, we were interested in examining whether self-assessment accuracy of correct boxes predicted absolute metacomprehension accuracy. The results of the multilevel regression analysis of absolute metacomprehension accuracy for both facts and relations (separately) as a function of self-assessment accuracy of correct boxes are displayed in Table 4. The model in the case of both facts and relations is statistically significant. Nonetheless, self-assessment accuracy of correct boxes explained only 2.9% ($R^2 = 0.029$) of the variance of absolute metacomprehension accuracy for facts and 6.8% ($R^2 = 0.068$) for that of relations. Therefore, the effect seems to be small despite the statistical significance of the model.

³ Accuracy percentages are determined from the values of Table 3 by the formula $(1 - \text{value} / \text{max_possible_value}) * 100$. Example: $(1 - 0.18 / 4) * 100 = 96\%$.

Diagnosticity and utilization of diagram cues (H2)

Means and standard deviations for the diagnosticity and utilization of diagram cues per diagram condition are listed in Table 5. Logically, since the intervention targeted relations, none of the cues were diagnostic for test scores on the factual questions in either condition ($r \leq .09$). Regarding test scores on causal relations, for both diagramming conditions, the number of correct boxes was the most diagnostic cue ($r = .54$), followed by the equally diagnostic number of omissions and completed boxes ($|r| = 0.34$). Commission errors for relations were the least diagnostic cues with the weakest Pearson correlations ($|r| \leq 0.21$). There were no significant differences ($p > .060$) in cue diagnosticity for any of the cues between diagramming conditions.

In terms of cue utilization, the pattern was the same for facts and relations in both conditions. Omissions and completed boxes—which were relatively diagnostic—were the most utilized cues ($|r| \geq 0.42$), followed by correct boxes, which were the most diagnostic cue ($|r| \geq 0.39$). Commission errors, which were the least diagnostic cue, were also the least utilized cue ($r \leq .18$). The utilization of commission errors in the self-assessment condition was negligible ($r \leq .06$).

We hypothesized that students who self-assessed their diagrams would utilize the diagnostic cues more when monitoring text learning compared to students who only completed the diagrams. However, commission errors for relations (which was the least diagnostic cue) was the only cue showing significant differences between diagramming conditions, $t(265) = -2.466, p = .014$, Cohen's $d = -0.302$. Students in the self-assessment group utilized the commission errors cue for relations significantly less ($\gamma = -0.02$) than their counterparts who completed the diagram without self-assessment ($\gamma = 0.18$).

Metacomprehension accuracy (H3)

We hypothesized that students' metacomprehension accuracy (absolute and relative) would be higher in the DC+SA group compared to the DC and C groups, and higher in the DC group than in the C group. There were significant differences among conditions regarding relative metacomprehension accuracy for relations, $F(3, 384) = 4.175, p = .006, \eta_p^2 = 0.032$. Pairwise comparisons with Bonferroni adjustments showed that students in the DC and DC+SA conditions had indeed significantly higher relative metacomprehension accuracy than those in the control condition, $p = .007$ and $p = .040$ respectively. However, contrary to our expectations, there were no significant differences between the two diagramming conditions ($p = 1.000$). There were also no significant differences among conditions in terms of relative metacomprehension accuracy for facts, $F(3, 390) = 0.33, p = .804, \eta_p^2 < 0.003$.

Table 6 Parameter estimates from the multilevel analysis of absolute metacomprehension accuracy for facts and relations

Effects	<i>B</i>	<i>SE B</i>	<i>p</i>
Facts			
Control vs. Diagramming only	-0.136	0.085	0.108
Control vs. Diagramming+self-assessment	-0.071	0.086	0.405
Diagramming only vs. + self-assessment	0.064	0.077	0.402
Relations			
Control vs. Diagramming only	-0.209	0.096	0.030
Control vs. Diagramming+self-assessment	-0.245	0.087	0.005
Diagramming only vs. + self-assessment	-0.036	0.061	0.554

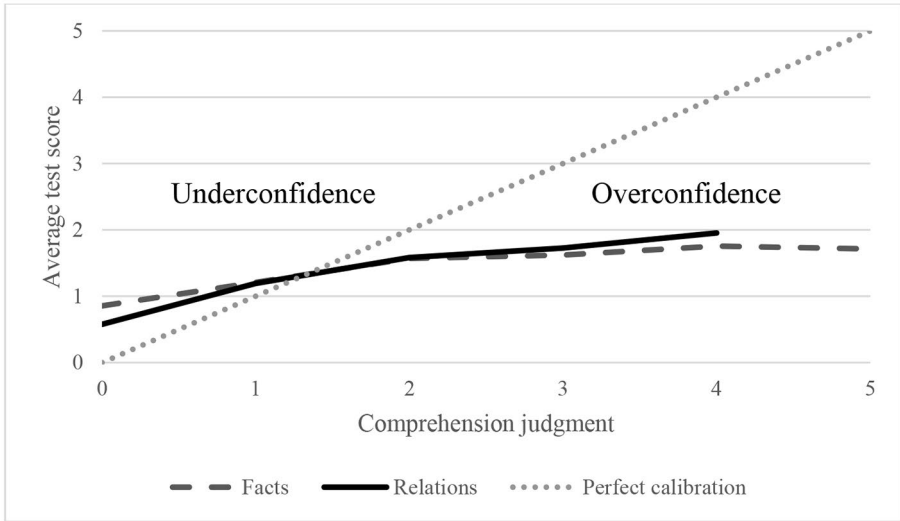


Fig. 4 Calibration curves of comprehension judgment bias for facts and causal relations

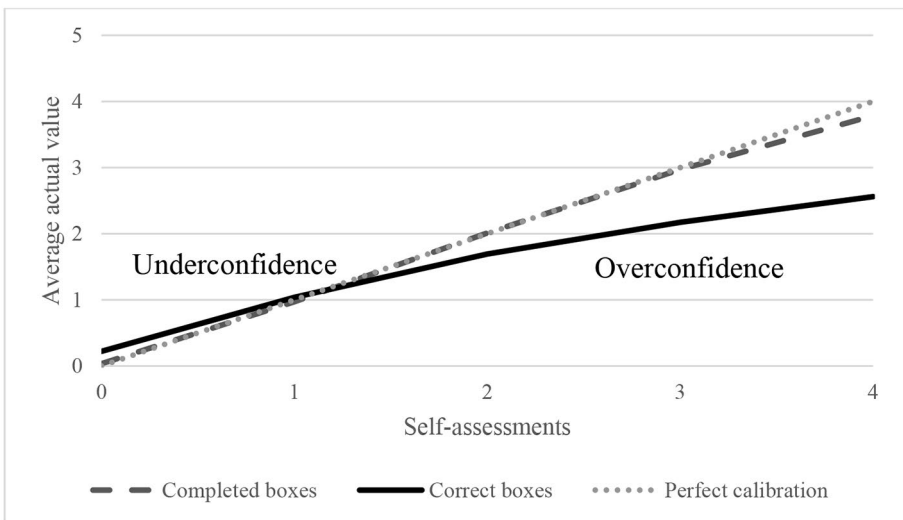


Fig. 5 Self-assessment calibration curves (bias) for completed and correct boxes

Analogous results were obtained for absolute metacomprehension accuracy (see Table 6) (see *Analyses*). For causal relations, both diagramming conditions led to significantly higher accuracy (two-tailed $p < .030$) than the control condition, but the two diagramming conditions did not differ (two-tailed $p = .554$). For facts, no pairwise comparisons showed significant differences (two-tailed $p > .108$).

On average, overconfidence (i.e., positive values of absolute metacomprehension accuracy, see Table 3) was found in all conditions for both facts and causal relations. The calibration curves for the bias in comprehension judgments and self-assessments are presented in Figs. 4 and 5 respectively. In both charts, the underconfidence and overconfidence areas are presented respectively to the left and right of the perfect calibration line. Both show the overconfidence tendency, and that overconfidence tends to grow as we get closer to the top of the scale.

Supplementary analyses

Even though the analyses provided here do not form part of the hypotheses, they are added for comparability with previous studies of the diagram scaffold and for completeness (e.g., for meta-analyses purposes). No significant differences were found among conditions in comprehension judgments for relations, $F(3, 2521)=1.651$, $p=.176$, $\eta_p^2 = 0.002$. Conversely, significant differences were found among conditions in terms of comprehension judgments for facts, $F(3, 2515) = 10.449$, $p < 0.001$, $\eta_p^2 = 0.012$. Students in the C condition gave significantly higher comprehension judgments for facts than their DC ($p < 0.001$) or DC+SA ($p = .001$) counterparts.

There were statistically significant differences among conditions for restudy selections, $F(3, 418)=6.447$, $p < 0.001$, $\eta_p^2 = 0.044$. Pairwise comparisons showed that students in the C condition selected significantly less texts for restudy than their counterparts in the DC ($p = .003$) and DC+SA ($p < 0.001$) groups.

Regulation

There were no statistically significant differences among conditions in terms of metacomprehension-based regulation for relations, $F(3, 368)=2.411$, $p=.067$, $\eta_p^2 = 0.019$ or metacomprehension-based regulation for facts, $F(3, 367) = 2.214$, $p = .086$, $\eta_p^2 = 0.018$. On the other hand, there were no significant differences either in comprehension-based regulation for relations, $F(3, 372) = 0.878$, $p = .542$, $\eta_p^2 = 0.007$; or comprehension-based regulation for facts, $F(3, 375)=0.346$, $p = .792$, $\eta_p^2 = 0.003$.

Reading comprehension

There were no significant differences among conditions in terms of test scores for relations, $F(3, 2538)=1.189$, $p=.130$, $\eta_p^2 = 0.002$. However, there were statistically significant differences among conditions for test scores for facts, $F(3, 2540) = 4.289$, $p = .005$, $\eta_p^2 = 0.005$, with students in the control condition scoring significantly higher than those in the diagramming-only condition ($p=.006$), and no other significant difference ($p > .165$).

Discussion

Metacomprehension accuracy and self-assessments

Contrary to our expectations, diagramming students with and without self-assessment did not differ significantly in their relative or absolute metacomprehension accuracy for causal relations. In other words, the self-assessment scaffold did not produce the intended theorized results. The only difference caused by prompting self-assessments was a significant decrease in the utilization of the commission errors cue (a cue that was found to have a low diagnostic value), to the extent that the utilization coefficient for that cue was negligible (i.e., not significantly different from zero).

Concurrently, comprehension judgments for causal relations correlated strongly with the self-assessment for correct boxes, which are the opposite of commission errors. Looking only at the DC+SA condition, one could be tempted to say that the scaffolding prompt for self-assessing correct boxes focused students' attention on correct boxes, the most diagnostic cue, in detriment of commission errors, the least diagnostic cue. However, that would have also caused differences in the utilization of the correct boxes cue between diagramming conditions (due to the difference between receiving the scaffolding prompt or not), which was not the case. It seems therefore more plausible to conclude that students implicitly and automatically self-assess their diagram answers, even without explicit instruction to do so. Such behavior could correspond to a natural process of self-monitoring during task completion (Ramdass & Zimmerman, 2008). The complete overlooking of the commission errors cue might be explained then by students implicitly monitoring all the diagram cues, but when prompted to self-assess correct boxes, the effect is not that they increase the attention they are already paying to it (or at least not significantly), but that the opposite or contrasting cue (i.e., commission errors) is overshadowed.

In line with previous studies (e.g., van de Pol et al., 2019; van Loon et al., 2014), diagramming proved again to be an effective scaffold to improve students' relative metacomprehension accuracy for causal relations when learning from expository texts. Such improvement was, however, less sizeable (on average up to a Goodman-Kruskal gamma of 0.36) than previous best reported results (on average up to 0.56). In any case, students in both diagramming conditions were roughly, on average, three times more accurate in terms of relative accuracy than those in the control condition. This enhancement is also visible in the contrast with relative accuracy for facts (not targeted by the intervention), where values did not differ significantly across conditions. The obtained relative accuracy is in between best and worst average results previously reported for other scaffolds in the form of generative interventions such as listing keywords (0.11–0.7; de Bruin et al., 2011; Thiede et al., 2003) and concept-mapping (0.35–0.67; Redford et al., 2012; Thiede et al., 2010); while lower than those reported for summarizing (0.55–0.60; Thiede et al., 2010; Thiede & Anderson, 2003).

Similarly, diagramming also increased absolute metacomprehension accuracy significantly. This increase, however, was less substantial (1.09 times better than the control condition) than the increase in relative accuracy (3.27 times better than the control condition). Noteworthy, students' monitoring was much more accurate in absolute than in relative terms; concretely, 6.27 times, 2.06 times, and 2.35 times respectively in the C, DC and DC+SA conditions. From this comparison across conditions we can also conclude that, although the gap between relative and absolute metacomprehension accuracy is still con-

siderable, diagramming contributed to reduce the gap. Adding to the literature contrasting the two types of accuracy, our results corroborate that they are statistically independent and show that average students' metacomprehension skills differ significantly from low relative accuracy to high absolute accuracy. Such a gap between students' skills concerning the two types of accuracy might originate in or relate to the nature and interpretation of each type. It comes to say that for students it is much more difficult to discriminate their comprehension judgments among different texts than to judge each text separately as it would correspond to absolute metacomprehension accuracy. A possible reason for the smaller effect on absolute than relative metacomprehension accuracy therefore is that absolute accuracy was overall higher and accordingly, less room was left for differences between conditions.

The overconfidence effect for absolute accuracy (Gigerenzer et al., 1991; Koriat, 2012; Labuhn et al., 2010) was once again replicated. Overconfidence seemed to account for the gap between high and perfect absolute metacomprehension accuracy. Similarly, overconfidence was the norm for self-assessments, but still their accuracy aligns with early findings in the metacomprehension literature that students can judge reasonably well the correctness of their answers, and with higher accuracy than their comprehension judgments (Glenberg & Epstein, 1985; Maki et al., 1994).

Cue diagnosticity and utilization

Comparing our findings to previous research (van de Pol et al., 2019, 2020; van Loon et al., 2014) in terms of cue diagnosticity and utilization, we observe that the patterns remain the same. The pattern of diagnosticity for diagram cues stably shows that the number of correctly completed boxes is the most diagnostic cue, while number of commission errors remains the least diagnostic. Analogously, the pattern of cue utilization remains that omission errors and completed boxes are the most utilized cues and commission errors the least, the latter being sometimes even negligible. Taken together, although the number of correct boxes is the most diagnostic diagram cue, it does not seem to be the most utilized by the students. Contrary to our hypothesis, the scaffold of prompting self-assessments did not translate into increased utilization of the number-of-correct-responses cue, maybe, as suspected above, because they already implicitly self-assess their answers as part of the task.

Regulation and reading comprehension

In concordance with previous diagramming studies (e.g., van de Pol et al., 2019; van Loon et al., 2014), a strong negative gamma correlation was found for metacomprehension-based regulation (i.e., association between comprehension judgments and restudy selections) both for facts and causal relations, with learners being more likely to select those texts for restudy that they judged their learning to be poor on. These results add to the evidence supporting the theoretical if-metacognitive-monitoring-then-metacognitive-control contingency (Winne, 2010) in the metacomprehension paradigm (Thiede et al., 2003). In other words, students' restudy selections (i.e., metacognitive control) are markedly influenced by their comprehension judgments (i.e., metacognitive monitoring).

Regarding comprehension-based regulation (i.e., association between restudy selections and reading comprehension), no significant difference was found between conditions. Nevertheless, the results for the control and both diagramming conditions are numerically quite

close to those of Van de Pol et al. (2019) in their control and diagram completion conditions, hinting at a stable effect.

Metacomprehension-based regulation, which accounts for the monitoring-control relationship, was over two times more effective for causal relations and over four times more effective for factual information than comprehension-based regulation, which accounts for a control-performance relationship. Although in different proportions, these findings replicate previous ones by Van de Pol et al. (2019) signaling a much stronger connection between monitoring and control (or regulation) than between control and performance. This is a quantified reflection of the consequences of low monitoring accuracy for students' academic performance. A direct actionable consequence is thus that improving the accuracy of comprehension judgments (i.e., monitoring) would automatically enhance that of comprehension-based regulation.

Regarding test scores, the results were different compared to previous studies that used the same materials. For causal relations, Van Loon et al. (2014) found that delayed-diagramming students significantly, and with a large effect size of $\eta_p^2 = 0.14$ (Cohen, 1988), outperformed control students in the test. Conversely, neither Van de Pol et al. (2019) nor us in this study found a significant difference for test scores of causal relations.

Limitations

Next, some potential limitations of the study are discussed. Pre-structured diagrams were provided to the students instead of asking them to create their own diagrams or concept maps. While the idea is to reduce the extraneous load of the intervention, as concept maps have been shown to require more training for their optimal effectiveness (Farrokhnia et al., 2019), it can also happen that the student, even knowing the causal relations perfectly, struggles to fit their causal relation structure into the diagram structure (e.g., sometimes they put together in a box what in the model answers belongs to two separate boxes). This situation can create some undesired and artificial misalignment in the cues. However, this potential limitation would apply equally to both diagramming conditions, thereby not affecting the relative comparison among them.

The results of the study, as pertains to the diagramming paradigm in general, correspond to the reading comprehension of causal relations. Different results might be obtained for example in the context of reading comprehension of declarative, conceptual, or procedural knowledge.

Furthermore, in the diagramming paradigm, regulation or control is limited to the restudy selections, but the repertoire of learning strategies is much wider (e.g., help-seeking and consulting other sources). Although regulation was not a central aspect of this study, it is important to consider it when analyzing the diagramming paradigm holistically in all the learning facets it represents (e.g., reading, generative learning, monitoring, and control).

Implications

The study has predominantly implications of theoretical and practical character since, methodologically, we followed the diagramming paradigm tradition, including materials and operationalizations, in order to maximize comparability with previous studies and results.

The study provided reinforcement for the theorized strong relationship between metacognitive monitoring and metacognitive control, which is more often assumed than shown empirically. The problematic of students' low relative metacomprehension accuracy is well-known in the metacognition literature and attested by reviews of tens of studies (e.g., Maki 1998; Prinz et al., 2020). This research showed that, conversely, students' absolute metacomprehension accuracy is high (six times higher than relative accuracy in the control condition, and two times higher in the diagramming conditions). Although absolute and relative accuracy are statistically independent from one another, an interesting implication of the diagram scaffold was that the gap between both types of accuracy was reduced.

Even more accurate than absolute metacomprehension were the self-assessments. Although mixed results have been found concerning self-assessment accuracy in the literature for different tasks and purposes, in the metacomprehension paradigm, of which diagramming is a particularization, their high accuracy has been shown to be much more stable, and our results align with those previous ones. An implication is therefore that teachers could pedagogically use students' self-assessments in tasks comparable to the metacomprehension paradigm and expect them to be reasonably accurate.

Last but not least, a practical implication is that teachers willing to help students to increase both their relative and absolute metacomprehension accuracy—and thereby their self-regulatory skills—of expository texts containing causal relations, can have them complete causal relations diagrams about the texts, without explicitly asking students to self-assess their diagrams; they seem to do that implicitly anyway.

Future Research

Future research should explore other scaffolds to focus students' attention more on the diagnostic cues when making comprehension judgments. Nudging theory, originally from behavioral economics but increasingly applied to education (Damgaard & Nielsen, 2018), offers a framework to influence behavior while keeping autonomy. An interesting follow-up research question, benefiting from cross-fertilization among disciplines, is then how to nudge students towards using more diagnostic cues.

Next to inappropriate cue utilization, inaccurate metacognitive monitoring can also result from an overload of cognitive resources (Butler & Winne, 1995). Theoretically, this warrants exploration in the light of Easterbrook's (1959) influential cue utilization theory, which explained a performance reduction under too high arousal via a sharpened selectivity in the attention focus, causing a narrowing in the number of cues attended to. Methodologically, these aspects are of increasingly accessible exploration via wearable sensors that enable measuring arousal in a continuous, unobtrusive, objective, and real-time fashion. We expect such avenues of research to produce valuable insights on metacomprehension accuracy shortcomings from a different, scarcely explored angle. Such approaches have already been undertaken by means of self-reported arousal (e.g., Prinz et al., 2019).

Opportunities for better understanding cue-utilization are also evident with eye-tracking research by identifying the diagram areas that students look at, both by fixations on specific boxes and by saccades while transitioning between boxes.

Nudging theory, arousal theory, wearable sensors, and eye-tracking glasses are just examples from a fertile ground of multidisciplinary approaches that could leverage theo-

retical and technological advances to further improve the already effective diagramming scaffold for raising metacomprehension accuracy.

Appendix A

Text “The Benefits of Music”.

Learning to play an instrument can have many benefits. This is because in order to read a sheet of music and to play notes one has to use various areas of the brain, such as the amygdala and the hippocampus. Canadian research findings showed that as a result of using these different brain areas, 12-year-old children who learned to play music achieved higher scores on an IQ test. In addition, using the brain while learning music can improve spatial skills, which are useful for solving math tasks, such as fractions. Furthermore, music can help with retrieving memories. For example, an old song can remind you of something that happened a long time ago. Patients with Alzheimer’s disease can, for example, also be supported to recall memories through music. This was shown in research by the scientist Dr. Polk. He examined an elderly woman with Alzheimer’s, who turned out to recognize all the pieces of music she had learned in the past.

Text “Renovation of Concrete Buildings”.

Concrete is still a widely-used building material. For almost seventy years, many large buildings have been made of concrete. Today, concrete usually must be tested to indicate that it is a suitable material to build large buildings. When building with concrete, it is important to consider the consequences of the heating system that will be used in the building. Central heating can dry out concrete. When concrete dries, it shrinks. This means that concrete buildings tend to shrink and thus become smaller than they originally were. Because of this shrinkage, elevators in concrete buildings often get stuck at some point. That is why it is often necessary to make elevators smaller and to shorten the elevator shafts. In 1952, the so-called Mazzo building was made of concrete panels. Due to the effect of central heating, the Mazzo building shrank by 12 cm. Mr. Nicolas, an elevator mechanic, explained that extensive renovation work was needed to adjust the elevators of the Mazzo building.

Appendix B

SPSS syntax for relative metacomprehension accuracy ANCOVA analysis with dyslexia as a covariate. The ANCOVAs of the other variables were carried out similarly.

```
UNIANOVA RelativeMetacomprehensionAccuracy BY Condition Dyslexia.
/METHOD=SSTYPE(3).
/INTERCEPT=INCLUDE.
/POSTHOC=Condition(BONFERRONI).
/EMMEANS=TABLES(Condition) COMPARE ADJ(BONFERRONI).
/EMMEANS=TABLES(Dyslexia) COMPARE ADJ(BONFERRONI).
/PRINT ETASQ DESCRIPTIVE PARAMETER.
/CRITERIA=ALPHA(0.05).
/DESIGN=Condition Dyslexia.
```

Acknowledgements This research was partially funded by the Netherlands Initiative for Education Research (NRO; Interlinked Research Projects; grant number 40.5.18300.024 awarded to the last author and the second author). In addition, during the realization of part of this research, the second author was funded by a Veni grant (number: 451-16-012) from the Netherlands Organization for Scientific Research (NWO) awarded to the second author. We would also like to thank all students who helped to collect the data and Aniek Uittenboogaard, Virrie van Zutven, Desar Dauti, and Jonne Bloem for their assistance with coding the students' answers and interrater reliability.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

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

References

- Andrade, H. L. (2019). A critical review of research on student self-assessment. *Frontiers in Education, 4*. <https://doi.org/10.3389/educ.2019.00087>
- Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods, 24*(1), 1–19. <https://doi.org/10.1037/met0000195>.
- Balzer, W. K., Doherty, M. E., & O'Connor, R. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin, 106*(3), 410–433. <https://doi.org/10.1037/0033-2909.106.3.410>.
- Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444. <https://doi.org/10.1146/annurev-psych-113011-143823>.
- Björkman, M. (1994). Internal cue theory: Calibration and resolution of confidence in general knowledge. *Organizational Behavior and Human Decision Processes, 58*(3), 386–405. <https://doi.org/10.1006/obhd.1994.1043>.
- Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education, 18*, 529–549. <https://doi.org/10.1007/BF00138746>.
- Brown, G. T. L., & Harris, L. R. (2014). The future of self-assessment in classroom practice: Reframing self-assessment as a core competency. *Frontline Learning Research, 2*(1), 22–30. <https://doi.org/10.14786/flr.v2i1.24>.
- Brunswick, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). University of California Press.
- Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. <https://doi.org/10.2307/1170684>.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. <https://doi.org/10.4324/9780203771587>
- Damgaard, M. T., & Nielsen, H. S. (2018). Nudging in education. *Economics of Education Review, 64*, 313–342. <https://doi.org/10.1016/j.econedurev.2018.03.008>.
- de Bruin, A. B. H., Thiede, K. W., Camp, G., & Redford, J. S. (2011). Generating keywords improves meta-comprehension 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>.
- de Bruin, A. B. H., Dunlosky, J., & Cavalcanti, R. B. (2017). Monitoring and regulation of learning in medical education: The need for predictive cues. *Medical Education, 51*(6), 575–584. <https://doi.org/10.1111/medu.13267>.
- Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction, 24*(1), 4–14. <https://doi.org/10.1016/j.learninstruc.2012.06.001>.

- Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, 16(4), 228–232. <https://doi.org/10.1111/j.1467-8721.2007.00509.x>.
- Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction*, 24(1), 58–61. <https://doi.org/10.1016/j.learninstruc.2012.05.002>.
- 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., Badali, S., Rivers, M. L., & Rawson, K. A. (2020). The role of effort in understanding educational achievement: Objective effort as an explanatory construct versus effort as a student perception. *Educational Psychology Review*, 1–13. <https://doi.org/10.1007/s10648-020-09577-3>.
- Easterbrook, J. A. (1959). The effect of emotion on cue utilization and the organization of behavior. *Psychological Review*, 66(3), 183–201. <https://doi.org/10.1037/h0047707>.
- Farrokhnia, M., Pijera-Díaz, H. J., Noroozi, O., & Hatami, J. (2019). Computer-supported collaborative concept mapping: The effects of different instructional designs on conceptual understanding and knowledge co-construction. *Computers & Education*, 142(December), <https://doi.org/10.1016/j.compedu.2019.103640>.
- Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, 28, 717–741. <https://doi.org/10.1007/s10648-015-9348-9>.
- Gale, J. (1984). Self-assessment and self-remediation strategies. In E. S. Henderson, & M. B. Nathenson (Eds.), *Independent Learning in Higher Education* (pp. 98–140). Educational Technology Publications.
- Gigerenzer, G., Hoffrage, U., & Kleinbülting, H. (1991). Probabilistic mental models: A brunswikian theory of confidence. *Psychological Review*, 98(4), 506–528. <https://doi.org/10.1037/0033-295X.98.4.506>.
- Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning Memory and Cognition*, 11(4), 702–718. <https://doi.org/10.1037/0278-7393.11.1.4.702>.
- Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition*, 15(1), 84–93. <https://doi.org/10.3758/BF03197714>.
- Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition*, 10(6), 597–602. <https://doi.org/10.3758/BF03202442>.
- Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General*, 116(2), 119–136. <https://doi.org/10.1037/0096-3445.116.2.119>.
- Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268), 732–764. <https://doi.org/10.1080/01621459.1954.10501231>.
- Graesser, A. C. (2007). An introduction to strategic reading comprehension. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 3–26). Erlbaum.
- Händel, M., de Bruin, A. B. H., & Dresel, M. (2020). Individual differences in local and global metacognitive judgments. *Metacognition and Learning*, 15(1), 51–75. <https://doi.org/10.1007/s11409-020-09220-0>.
- Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, 56(4), 208–216. <https://doi.org/10.1037/h0022263>.
- Huff, J. D., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition and Learning*, 4(2), 161–176. <https://doi.org/10.1007/s11409-009-9042-8>.
- Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology*, 74(4), 657–690. <https://doi.org/10.1037/0021-9010.74.4.657>.
- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126(4), 349–370. <https://doi.org/10.1037/0096-3445.126.4.349>.
- Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, 119(1), 80–113. <https://doi.org/10.1037/a0025648>.
- Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2), 107–118. <https://doi.org/10.1037/0278-7393.6.2.107>.
- Kostons, D., van Gog, T., & Paas, F. (2012). Training self-assessment and task-selection skills: A cognitive approach to improving self-regulated learning. *Learning and Instruction*, 22(2), 121–132. <https://doi.org/10.1016/j.learninstruc.2011.08.004>.
- Labuhn, A. S., Zimmerman, B. J., & Hasselhorn, M. (2010). Enhancing students' self-regulation and mathematics performance: The influence of feedback and self-evaluative standards. *Metacognition and Learning*, 5(2), 173–194. <https://doi.org/10.1007/s11409-010-9056-2>.

- Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 131–158). Routledge. <https://doi.org/10.4324/9781410602350-13>
- Maki, R. H., Jonas, D., & Kallod, M. (1994). The relationship between comprehension and metacomprehension ability. *Psychonomic Bulletin & Review*, 1(1), 126–129. <https://doi.org/10.3758/BF03200769>.
- Muthén, B., & Muthén, L. (2017). Mplus. In van der W. J. Linden (Ed.), *Handbook of Item Response Theory* (pp. 507–518). Chapman and Hall/CRC. <https://doi.org/10.1201/9781315117430-28>.
- Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, 95(1), 109–133. <https://doi.org/10.1037//0033-2909.95.1.109>.
- Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect. *Psychological Science*, 2(4), 267–271. <https://doi.org/10.1111/j.1467-9280.1991.tb00147.x>.
- Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, 26, 125–173. [https://doi.org/10.1016/S0079-7421\(08\)60053-5](https://doi.org/10.1016/S0079-7421(08)60053-5).
- Nystedt, L., & Magnusson, D. (1973). Cue relevance and feedback in a clinical prediction task. *Organizational Behavior and Human Performance*, 9(1), 100–109. [https://doi.org/10.1016/0030-5073\(73\)90039-1](https://doi.org/10.1016/0030-5073(73)90039-1).
- Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, 48(1), 85–112. <https://doi.org/10.1016/j.jsp.2009.09.002>.
- Pierce, B. H., & Smith, S. M. (2001). The postdiction superiority effect in metacomprehension of text. *Memory & Cognition*, 29(1), 62–67. <https://doi.org/10.3758/BF03195741>.
- Pieschl, S. (2009). Metacognitive calibration—an extended conceptualization and potential applications. *Metacognition and Learning*, 4(1), 3–31. <https://doi.org/10.1007/s11409-008-9030-4>.
- Prinz, A., Bergmann, V., & Wittwer, J. (2019). Happy but overconfident: Positive affect leads to inaccurate metacomprehension. *Cognition and Emotion*, 33(3), 606–615. <https://doi.org/10.1080/02699931.2018.1472553>.
- Prinz, A., Golke, S., & Wittwer, J. (2020). 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>.
- Ramdass, D., & Zimmerman, B. J. (2008). Effects of self-correction strategy training on middle school students' self-efficacy, self-evaluation, and mathematics division learning. *Journal of Advanced Academics*, 20(1), 18–41. <https://doi.org/10.4219/jaa-2008-869>.
- 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>.
- Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin*, 137(1), 131–148. <https://doi.org/10.1037/a0021705>.
- Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review*, 12, 437–475. <https://doi.org/10.1023/A:1009084430926>.
- 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., Griffin, T. D., Wiley, J., & Redford, J. S. (2009). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 85–106). Routledge. <https://doi.org/10.4324/9780203876428>
- 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>.
- 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 Loon, M. H., van Gog, T., Braumann, S., & de Bruin, A. B. H. (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 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(September), 143–154. <https://doi.org/10.1016/j.actpsy.2014.06.007>.

- Walczyk, J. J., & Hall, V. C. (1989). Effects of examples and embedded questions on the accuracy of comprehension self-assessments. *Journal of Educational Psychology, 81*(3), 435–437. <https://doi.org/10.1037/0022-0663.81.3.435>.
- Weingardt, K. R., Leonesio, R. J., & Loftus, E. F. (1994). Viewing eyewitness research from a metacognitive perspective. In J. Metcalfe, & A. P. Shimamura (Eds.), *Metacognition, knowing about knowing* (pp. 157–184). MIT.
- Wiley, J., Griffin, T. D., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *Journal of General Psychology, 132*(4), 408–428. <https://doi.org/10.3200/GENP.132.4.408-428>.
- Williams, J. P. (2007). Literacy in the curriculum: Integrating text structure and content area instruction. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 199–219). Erlbaum.
- Winne, P. H. (2010). Improving measurements of self-regulated learning. *Educational Psychologist, 45*(4), 267–276. <https://doi.org/10.1080/00461520.2010.517150>.
- Yan, Z., Lao, H., Panadero, E., Fernández-Castilla, B., Yang, L., & Yang, M. (2022). Effects of self-assessment and peer-assessment interventions on academic performance: A meta-analysis. *Educational Research Review, 37*, <https://doi.org/10.1016/j.edurev.2022.100484>.
- York, K. M., Doherty, M. E., & Kamouri, J. (1987). The influence of cue unreliability on judgment in a multiple cue probability learning task. *Organizational Behavior and Human Decision Processes, 39*(3), 303–317. [https://doi.org/10.1016/0749-5978\(87\)90026-4](https://doi.org/10.1016/0749-5978(87)90026-4).

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.