Instructional Science
https://doi.org/10.1007/s11251-020-09521-6

ORIGINAL RESEARCH

Is drawing after learning effective for metacognitive monitoring only when supported by spatial scaffolds?

Julia Kollmer[1] · Katrin Schleinschok[2,3] · Katharina Scheiter[2,3] · Alexander Eitel[4]

Received: 23 July 2019 / Accepted: 25 June 2020

Abstract
In this study, we investigated whether drawing after learning supports metacognitive monitoring especially when students are supported in their drawing efforts. Therefore, eighty-eight participants were randomly assigned to one of three experimental groups. They were asked to learn from a text comprising five paragraphs about the formation of auroras. After reading each of the five paragraphs, one group had to mentally imagine the contents (control group), a second group had to draw from scratch, and a third group had to draw with the help of spatial scaffolds. All participants provided judgments of learning (JOL) for each paragraph, and took a knowledge test afterwards. Results revealed that students who drew, both with and without scaffold, monitored their learning more accurately on an absolute level. Even though there were no differences between the two drawing conditions for monitoring accuracy, JOLs were based on the actual drawing quality only when students drew with the help of spatial scaffolds. Results thus hint towards the potential of (scaffolded) drawing to support metacognitive monitoring. Reasons for why drawing with spatial scaffolds did not improve monitoring compared to drawing from scratch are discussed.

Keywords Self-regulated learning · Metacognition · Metacognitive monitoring · Calibration · Drawing · Learning from text

Introduction

With the advent of digital technology, the needs, as well as the opportunities, for autonomous learning outside of classroom settings have grown. Due to an increasing number of learning environments without the guidance of teachers, the ability to self-regulate one's

The original article has been corrected due to retrospective open access.

✉ Julia Kollmer
julia.kollmer@psychologie.uni-freiburg.de

[1] Department of Psychology, University of Freiburg, 79085 Freiburg, Germany

[2] Leibniz-Institut für Wissensmedien, Tübingen, Germany

[3] University of Tübingen, Tübingen, Germany

[4] Justus-Liebig University, Giessen, Germany

Published online: 10 July 2020

own learning becomes more relevant. "Knowing how to manage one's own learning activities has become, in short, an important survival tool" (Bjork et al. 2013, p. 418). Even if "survival" seems to be slightly overstated, it must be admitted that self-regulated learning (SRL) has become crucial for private and professional development. Thus, there is a need to enable learners to set learning goals, direct their learning behavior, and monitor whether their goals have been achieved. Briefly, it is important to enable learners to self-regulate their learning.

SRL requires learners, among other things, to know what they have already learned and understood (monitoring), and to adapt their study behavior accordingly (control). Therefore, they need to know which contents need further studying in order to reach the learning goals. Unfortunately, learners often do not have good insight into what they already know, which means that their metacognitive monitoring is inaccurate (e.g., Bjork et al. 2013). In consequence they need support. How such support may look like is the topic of the present study. In particular, the present study tests whether (supported) drawing can support metacognitive monitoring by using a classic paradigm to study SRL, which will be described in the following.

The classic paradigm to study metacognitive monitoring and regulation

In typical studies on metacognitive monitoring (e.g., Nelson and Dunlosky 1991; Thiede et al. 2003) students first have to learn something (e.g., from an expository text) and afterwards they provide *judgments of learning* (JOL), meaning that they indicate their level of confidence about how well they think they have understood and/or will be able to retrieve the instructional contents (e.g. on a scale from 0 to 100%). Sometimes students are provided with the opportunity to restudy after they gave their JOLs. Then they work on a knowledge test. This experimental procedure is known as the *metamemory paradigm* (de Bruin and van Gog 2012; Nelson and Narens 1990). Using this paradigm, researchers usually study the correspondence between students' judgments of their learning and their actual learning (based on their scores in the knowledge test). The latter correspondence is termed monitoring accuracy.

In most prior research, two types of monitoring accuracy are distinguished (cf. Serra and Metcalfe 2009). The first one is absolute monitoring accuracy, bias, or calibration (e.g., Alexander 2013; Dunlosky and Rawson 2012). It is calculated as the difference between a learner's JOL and his or her actual performance. Values near zero indicate only slight differences between estimated and actual performance, and therefore indicate good monitoring. Positive values indicate overconfidence in one's learning, whereas negative values indicate underconfidence. The higher a value is (regardless of the sign), the less accurate is the learner's estimation.

Secondly, relative monitoring accuracy, or resolution, indicates how well a learner is able to discriminate between well-learned and less well-learned materials (Lichtenstein and Fischhoff 1980; Nelson and Dunlosky 1991). Relative monitoring accuracy is commonly measured using a Goodman and Kruskal's gamma correlation between the participants' JOLs and their actual performance across several learning materials. To be able to calculate this measure, studies on relative monitoring accuracy typically have the same students learn multiple expository texts, often with unrelated topics (e.g., Thiede et al. 2003). High relative monitoring accuracy is indicated by a correlation near + 1.0, whereas correlations near zero indicate poor accuracy. Relative monitoring accuracy is particularly important

for control proceses in the context of academic learning, which often requires decisions whether a certain learning material has to be studied again or can be dropped.

In some studies, it was also investigated whether subsequent control behavior is successful. Regulation behavior is measured by providing students with the opportunity to restudy and by investigating whether students' JOLs correspond to subsequent study decisions (e.g., restudy time) as well as whether restudy time is associated with posttest performance (Pieger et al. 2016; Thiede and Dunlosky 1999). The idea is that accurate monitoring leads to more adequate restudy decisions (e.g., to close knowledge gaps) so that final performance should be better. Typically, monitoring accuracy is the main dependent variable in research on metacognition and learning. It is considered a prerequisite for successful SRL, because subsequent control decisions are assumed to be based on it. Hence, monitoring accuracy is also the main dependent variable here.

How to improve metacognitive monitoring

Previous research has shown that monitoring is often inaccurate—both on an absolute and on a relative level, especially when learning complex contents by means of expository texts (e.g., Eitel 2016). Specifically, students were found to be overconfident (e.g., Eitel 2016; Serra and Dunlosky 2010), meaning that their JOLs were higher than their actual performance (standardized on scale from 0 to 100). Moreover, students were often found to have low to medium resolution levels (e.g., Jaeger and Wiley 2014; Nelson and Dunlosky 1991; Rawson and Dunlosky 2002). Such inaccuracies in monitoring pose a major challenge for successful learning, because students may fail to adequately regulate their learning in turn. For instance, overconfidence can lead to underachievement (cf. Dunlosky and Rawson 2012) because students may prematurely believe that they have understood the contents well enough, and stop studying too early (cf. discrepancy reduction model; Thiede and Dunlosky 1999). Hence, the question that we try to answer here is how to improve SRL by improving monitoring accuracy.

To improve monitoring accuracy, it is important to first search for the reasons behind low monitoring accuracy. Low monitoring accuracy can be attributed to students' common beliefs and heuristics about learning and memory (e.g., Bjork et al. 2013; Serra and Metcalfe 2009). For instance, after (passively) reading text, students may base their JOLs on a superficial level of comprehension. They may fall prey to the *ease-of-processing heuristic* (e.g., Rawson and Dunlosky 2002), meaning that because it was easy to read through a text, its contents are judged as easy and as well retrievable at a later point in time (i.e., higher JOLs are provided). However, the ease of reading through a text is often not a reliable cue for later successful retrieval.

To improve calibration, students should provide JOLs based on whether they understood text on a deeper level of comprehension. Such a deeper level of comprehension is represented in a situation model according to the *construction-integration model* (van Dijk and Kintsch 1983). Comprehension on a situation-model level entails 'reading between the lines'; it means to construct a representation that goes beyond what was merely stated in the text. To achieve this kind of comprehension, students need to integrate ideas from the text with the help of their prior knowledge. Comprehension on a situation-model level is more resistant to forgetting than the two more shallow forms of text comprehension, namely the text base representing which propositions are directly stated in the text and the text surface representing text in a verbatim manner (Kintsch et al. 1990). Hence, JOLs that are based on situation model comprehension might reflect performance in later knowledge

tests more accurately (Redford et al. 2012). That is largely because of a reduced *foresight bias*, meaning less of an illusion that what is well-retrievable when providing JOLs (right after reading text) will be well-retrievable in later posttests (Koriat and Bjork 2006). Basing JOLs on the more shallow forms of text comprehension, text surface or text base, is likely related with inflated judgements (a stronger foresight bias), because substantial forgetting of these information takes place until the time of the posttest. Hence, posttest scores are likely to be lower than JOLs reflecting suboptimal calibration. Because, by contrast, less forgetting takes place on the situation-model level (cf. Kintsch et al. 1990), JOLs are likely to be less inflated, and thus calibration more accurate, when they are based on situation-model comprehension.

It follows from the previous paragraph that prompting students to engage in situation-model reasoning (rather than text-base or text-surface) prior to providing their JOLs may foster monitoring accuracy, specifically calibration. A commonly used method to achieve this is to have students perform a generative task (i.e., a task in which students have to generate words, sentences, drawings, etc.) after reading text and directly before providing the JOLs (e.g., Thiede et al. 2003; van Loon et al. 2014). Generative tasks such as drawing or concept mapping require situation-model reasoning, because they require constructing a 'new' representation in a different code. Specifically, they require both the transformation of concepts mentioned from the text into a visual format, and the integration of these concepts into a coherent visual display (Hilbert and Renkl 2008; van Meter and Garner 2005). Verbally expressed relations need to be transformed to spatial relations—a process that requires students to mentally represent and manipulate spatial information (i.e. spatial ability; Hegarty and Waller 2005). Through mapping or drawing, students receive implicit feedback, as the externalization helps them experience possible gaps in their knowledge. Accordingly, previous studies have found that both drawing and concept mapping support monitoring accuracy (Kostons and de Koning 2017; Redford et al. 2012; Schleinschok et al. 2017). Whether and how scaffolded drawing further supports monitoring accuracy will be in the focus of the present study.

How to improve drawing to further improve metacognitive monitoring

Previous research reveals preliminary evidence for drawing as a means to support metacognitive monitoring, both on an absolute (Kostons and de Koning 2017) and on a relative level (Schleinschok et al. 2017). Specifically, in the study of Schleinschok et al. (2017) students read five text sections about the formation of auroras; afterwards, they were either prompted to provide free-hand drawings on blank paper, or not. Students who drew after reading the text sections showed more accurate metacognitive monitoring.

There are two limitations to the study of Schleinschok et al. (2017) that motivated the present research. First, the free-hand drawing task was compared to performing no task; this produced time-on-task differences between the two conditions. To rule out differences in time-on-task as alternative explanation, the control condition would need to have an on-topic control task that is comparable in length to the drawing task. In the context of drawing, a fair control task is to have students mentally reactivate the contents that the students from the other condition have to draw (cf. Leutner et al. 2009). Hence, a mental reactivation prompt was used in the control condition in the present study.

Second, results from Schleinschok et al. (2017) revealed that drawing fostered relative but not absolute monitoring accuracy. One reason might be that the students in this study concentrated on filling their blank paper with details that were mentioned in the text (as many as

possible, regardless of how relevant) instead of focusing on producing a coherent image of the crucial elements to be understood. Hence, students might have based their JOLs on the fact they could draw as many elements as possible rather than on the actual drawing quality—that is, the accuracy of the drawing relative to the information provided in the text or to an expert's representation, or in other words the degree to which the drawings represent the linguistic information in a comprehensive and correct fashion (Scheiter et al. 2017). Especially the latter should reflect the situation-model reasoning that is diagnostic for good performance in a later knowledge test, leading to accurate JOLs. To put it differently, drawing as a metacognitive cue might not be diagnostic enough for learning outcomes when students need to draw from scratch. Students might need to be supported in their drawing efforts to monitor their learning more accurately, for instance, by receiving scaffolds for their drawings (e.g., a pre-drawn outline). In a similar vein, previous research by Leutner et al. (2009) showed that students focus their drawings more on crucial elements when they are provided with scaffolds that prespecify a spatial frame to which the crucial elements should just be added (supporting focused processing; Renkl 2014). The idea behind such spatial scaffolds or frames is that they reduce the degrees of freedom and therefore the number of decisions to be made when producing a drawing. Students do not need to think about where to start (in the middle or at the borders), whether they should draw all elements or just those that are considered being very crucial (e.g. because mentioned several times in the texts). The scaffolds directly prompt where to draw and (in part) what to draw. Hence, such scaffolds can be helpful to foster focused drawing (cf. Leutner et al. 2009; see also Schmidgall 2017). In previous research with open-book drawing, where the text is present when the drawing is produced (e.g., Leutner et al. 2009), such focused drawing led to better performance in subsequent knowledge tests. Here we will focus on the effects of closed-book drawing, where the drawing is produced after the text has been presented (cf. Schleinschok et al. 2017). In this situation, drawing should primarily act as feedback for how well one has understood the text contents rather than as direct learning aid. Thus, we expect (scaffolded) drawing to foster monitoring accuracy rather than direct learning gains.

The present research and hypotheses

We sought to test whether (scaffolded) drawing fosters absolute and relative monitoring accuracy by having students either mentally imagine the text contents after reading them or draw them from scratch (on blank paper) or draw them with the help of a spatial scaffold. The following two hypotheses were derived.

Monitoring accuracy hypothesis

We first hypothesized that students who draw (with and without scaffold) would monitor their learning and understanding more accurately on an absolute and relative level than students who are not asked to draw. Second, we hypothesized that especially when drawing with a spatial scaffold, students should have good monitoring accuracy.

Use-of-diagnostic-cue hypothesis

We hypothesized that especially after drawing with a spatial scaffold students would base their JOLs on whether they were able to draw a coherent image of the crucial elements. Hence, especially in this condition drawing quality should predict the JOLs. Conversely,

students who drew from scratch would base their judgments also on less diagnostic cues like, for instance, the number of drawn concepts, which could be misleading if many irrelevant details were drawn. Hence, in this condition drawing quality should predict the provided JOLs to a lesser degree.

Method

Participants and design

We calculated the a priori statistical power using G*Power (Faul et al. 2007). Because we expected a specific contrast to best explain the results, power calculations were based on a multiple regression analysis with two independent predictors (focal contrast, and one residual contrast), R^2 deviation from zero, and a medium-sized effect, $f^2 = 0.15$, with $\alpha = 0.05$, $1 - \beta = 0.80$. This power analysis revealed a minimum required sample size of 68. To make sure that the power is safely above 0.80, we sampled more participants than the calculated minimum, namely 90 students (79 female; $M = 22.5$ years, $SD = 3.1$) from two universities in the southern part of Germany. All participants received 10 Euro for their participation. Physics, chemistry, geology and meteorology students were not allowed to take part in the study to avoid that participants would possess too much prior knowledge about the learning content (formation of auroras). Participants were randomly assigned to one of three groups (control, drawing-from-scratch, drawing with scaffold). Data from two participants had to be excluded from the data analysis. One participant's fields of study were engineering and physics. The second participant was color-blind, which could have interfered with the color coding used in the material. Thus, there were 29 participants in the control group, 30 participants in the scaffolded-drawing group and 29 participants in the drawing-from-scratch group.

Materials

Instructional text

The instructional material was an expository text on the formation of auroras that described the processes of this phenomenon at a physical–chemical level (e.g. excitation of atmospheric molecules through electrons and protons). It was a slightly modified version of the text used by Schleinschok et al. (2017). It was developed at the Leibniz-Institut für Wissensmedien in Tübingen (Germany), and was pretested in several studies. The text comprised 1,311 words and consisted of five paragraphs, in which the formation of auroras was introduced step by step. The instructional material consisted of pure text and included no diagrams or other visualization (cf. sample paragraph in "Appendix").

Text difficulty was measured with the Flesch-Kincaid reading score (German version; Lenhard and Lenhard 2014). The score of 54 indicated that the text was relatively difficult to read. Reading scores between 50 and 60 correspond to the difficulty level of nonfictional literature. The text was divided into five paragraphs that were presented on separate pages. All paragraphs comprised a high degree of spatial information. For instance, the third paragraph, entitled 'the encounter of two magnetic fields', described the slowdown of the solar winds when they reach the earth's magnetic field. The solar winds stream around the obstacle. As a result, the earth's magnetic field gets clinched at one side and stretched

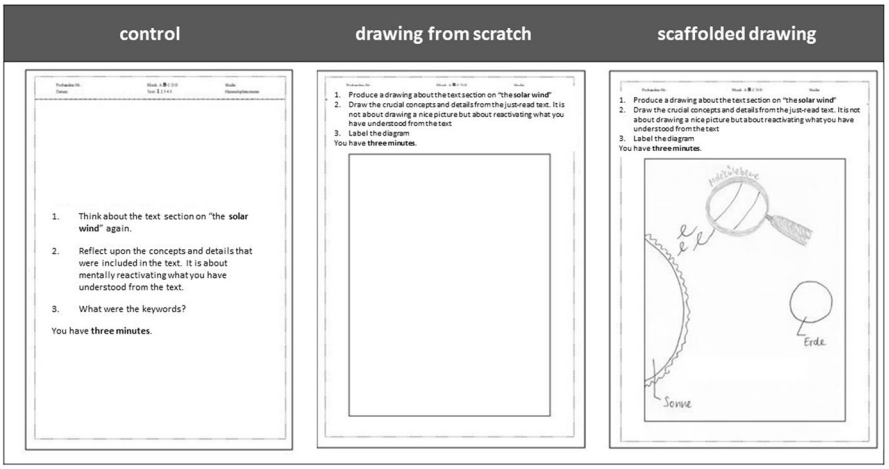


Fig. 1 Instructions the learners got after they read each paragraph of the text in the control condition (left), in the drawing-from-scratch (middle), and in the scaffolded-drawing condition (right)

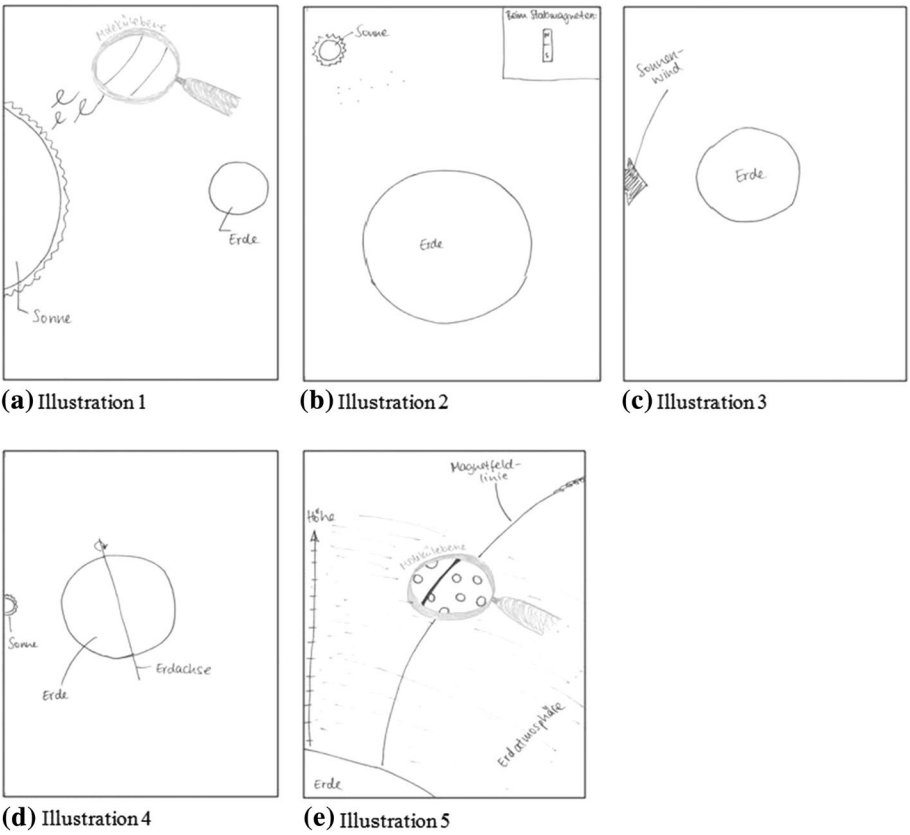
to a tail at the other side. Through the melted magnetic field lines in the tail, ions enter the earth's magnetic field and move towards the earth. The single paragraphs did not refer to previous paragraphs. Thus it was possible to understand the content of a paragraph without understanding the previous ones.

Experimental manipulation

Depending on the experimental condition, participants received different instructions after they finished reading each one of the five text sections. In all three conditions, the instructions comprised three aspects displayed in a list format on one page (see Fig. 1). The three aspects were the same in all three conditions.

The control group was mainly instructed to mentally reactivate the just-read contents. Specifically, they were told to "*1. Think about the text section on 'the solar wind' again. 2. Reflect upon the concepts and details that were included in the text. It is about mentally reactivating what you have understood from the text. 3. What were the keywords? You have three minutes*." Students in both the drawing-from-scratch and the scaffolded-drawing condition were instructed to "*1. Produce a drawing about the text section on 'the solar wind'. 2. Draw the crucial concepts and details from the just-read text. It is not about drawing a nice picture but about reactivating what you have understood from the text. 3. Label the diagram. You have three minutes."* Hence, the two drawing conditions did not differ from each other apart from the fact that students in the drawing-from-scratch condition drew on blank paper (more specifically: into an empty frame; see Fig. 1) whereas students in the scaffolded-drawing condition drew to complete missing links and concepts of provided illustrations. Figure 2 displays the provided illustrations for the scaffolded-drawing group (one illustration for each paragraph in chronological order).

The provided illustrations in the scaffolded-drawing group were specifically designed to not deliver details mentioned in text and to not prompt crucial contents. Instead, the illustrations primarily comprised labelled sketches of well-known elements (e.g. of the sun and the earth) and their relations to each other. These illustrations should primarily help students



(a) Illustration 1 **(b)** Illustration 2 **(c)** Illustration 3

(d) Illustration 4 **(e)** Illustration 5

Fig. 2 Spatial scaffolds for students' drawings for the five text paragraphs in the scaffolded-drawing condition. Students were prompted to draw the missing links

decide where to start drawing, and of what size a suitable drawing might be. But the illustrations did not depict the crucial elements and their relations from text by themselves. As the post-test primarily assessed understanding and recalling of the crucial elements and their relations from text, one can safely conclude that the illustrations in the scaffolded-drawing group did not directly provide the information required for the post-test. For instance, paragraphs 1 and 5 dealt with the chemical composition of the solar winds and the earths' atmosphere, respectively. The illustrations for these paragraphs contained, among others, a depicted magnifying glass (see Fig. 2a, e) so students knew an appropriate place where to start their part of the drawing about the chemical composition of the solar wind and the earths' atmosphere. All provided outlines were labeled to avoid errors in interpretation.

Measures

Control variables

We coded participants' answers to demographic questions about age, gender, school grades, and study courses as well as to six questions about their estimated current prior

knowledge in physics, particularly on the formation of auroras (e.g., how much do you know about solar winds?"; Scale from 0 = 'nothing' to 100 = 'very much').

We used ten items of the Paper Folding Test (Ekstrom et al. 1976) to measure participants' visuo-spatial abilities, because the learning task required visuo-spatial reasoning. During this test participants should imagine how a folded paper with punches looks like when it is unfolded, and choose the correct one of five alternative answers. For correct responses one point is awarded, for each error one point is subtracted (cf. Ekstrom et al. 1976).

Dependent variables

The dependent variables were (1) drawing quality, (2) judgments of learning (JOL) and (3) performance. These were used to calculate measures for absolute and relative monitoring accuracy, as well as to assess whether diagnostic cues were used for monitoring. (1) We assessed the quality of the drawings by identifying the crucial aspects of each paragraph and by coding whether they were correctly visualized. Participants received one point for each crucial aspect that they correctly drew, yielding 36 points in total. Two raters scored 25% of the drawing data. Inter-rater agreement regarding the drawing quality for each paragraph was sufficiently high (all ICCs > 0.84, for two-way random-effects model with measures of consistency) so that one of the two raters scored the remaining data. For each paragraph we calculated the percentage correct as measure for the drawing quality. (2) Participants provided JOLs for each of the five paragraphs; they were literally asked, "How confident are you that you are able to correctly answer questions regarding the paragraph 'The Solar Wind' you have just read?".

An 11-point rating scale was provided with the end points of 0 = 'not confident', and 100 = 'very confident'. (3) The post-test was intended to assess the actual performance. It was developed at the Leibniz-Institut für Wissensmedien in Tübingen (Germany) and was pretested in several studies. It contained eight verification items and one picture-labeling task per paragraph, yielding 40 verification items and five picture-labeling tasks in total. The verification items consisted both of statements that could be answered by merely recalling the respective information from the text on a propositional, text-base level (e.g. 'Plasma consists mostly of split atoms') and of statements that required an inference to be drawn to answer correctly (situation-model level). For instance, correctly responding to the statement "a strong activity of the sun can impair airplane travel" (without guessing) required not just reactivating the information from text that "the sun wind (the electrically charged particles) creates electromagnetic fields that can lead to malfunction of electronic and technological devices such as satellites" but also the drawing of an inference that airplanes are such electronic and technological devices so that airplane travel is potentially impaired. Another item that required the learners to draw inferences was for example "Auroras occur especially around the magnetic field poles because most of the magnetic field lines join there." In order to correctly reject this statement, the learners must have understood that the occurrence of auroras is not a question of where the magnetic field lines join in, but depends on which magnetic field lines carry electrically charged particles. For each item participants had to decide whether the statements were true or false. For each correct selection, they were awarded with one point; other responses (e.g., no selection, double selection, and incorrect selection) were coded as zero points. Points were summarized for each paragraph separately; they ranged from zero to eight. Test scores were adjusted for guessing probability by subtracting four points (half of maximum). Thus, 50

percent of the right answers were ruled out because there is a 0.5 probability of guessing the right answer by chance. Subsequently, the adjusted scores were converted to percentages, in order to match the JOL scale ranging from 0 to 100.

The picture-labeling task required students to find correct labels for relevant components of an illustration depicting the learning contents. As this was the first time the students saw illustrations about the crucial text contents they read before, they needed to retrieve and *map* the terms from text with the corresponding parts in the illustration. Thus, students needed to refer to their prior knowledge to be able to decide which of the depicted elements best match a remembered term from the previously read text, which (also) requires situation-model reasoning. There was one illustration for each paragraph containing 8–13 components to label. To code the participants' answers, the scoring scheme used in Schleinschok et al. (2017) study was modified, for each label we defined an extensive list of acceptable answers. Overall, the maximum score for picture-labeling tasks was 40 points; the minimum score was zero. Points were summarized for each paragraph separately and converted to percentages. Percentage scores for the verification items and picture-labeling tasks were averaged for each participant per paragraph resulting in five percentage scores for post-test performance per participant. We averaged across verification and picture-labeling scores primarily because (1) both of them assessed students' knowledge about the text contents, (2) both scores were substantially correlated ($r = 0.64$, $p < 0.001$) and did not differ significantly across conditions (both $p_s > 0.50$, see descriptive values in Table 1), and (3) most importantly, an unspecific (aggregated) performance score is better matched to the unspecific JOL assessment as it was generally asked about the confidence to "correctly answer questions regarding the paragraph XY" without referring to the type or format of the questions. Correlations between the percentage scores per each of the five paragraphs were sufficiently high, indicating an acceptable level of reliability for the whole test (Cronbach's alpha = 0.80).

Procedure

Data collection took place in 28 sessions with one up to nine participants. The experimenter was always the same person. The experimental manipulation was randomized at the level of sessions. All instructions and materials were presented as paper–pencil materials. The experimental procedure contained five steps (A–E) that were identical in time and sequence for all participants. The manipulation took place in part B. Starting the study, students were asked to fill out the demographic questionnaire and answer questions about their prior knowledge. Subsequently in part B, participants were instructed to carefully read and learn the text and after each paragraph, depending on condition, to either (1) think about the content of the paragraph (control), (2) draw a picture from scratch, or (3) draw a picture with the help of a spatial scaffold. Immediately after finishing the experimental task, students provided their JOL for the just read paragraph. While they worked on the tasks and gave the JOLs, participants did not have access to the paragraph. The same procedure (read paragraph, work on task, provide JOL) was repeated for all five paragraphs. To ensure that participants in all conditions were equally exposed to the learning material, the time to read and learn the paragraphs was fixed. For each paragraph participants had three to four minutes of study time (depending on the length of the paragraph), with an additional three minutes for the experimental task. Afterwards, the Paper Folding Test was administered (part C). In part D, participants were asked to select paragraphs for restudy. In fact, participants had no opportunity to restudy any of the paragraphs to not confound

Table 1 Means (SD) and empirical obtained minima and maxima for students' entry characteristics, overall JOLs and performance

	Min	Max	Control	Drawing from scratch	Scaffolded drawing	All subjects
Age	19	33	22.10 (3.00)	22.93 (3.24)	22.57 (3.12)	22.53 (3.10)
Visuo-spatial abiliies[a]	− 8	9	4.66 (2.87)	4.36 (2.74)	3.73 (4.20)	4.24 (3.34)
Latest school grades in physics[b]	4	1	2.52 (0.91)	2.59 (0.64)	2.70 (0.79)	2.61 (0.78)
Self-rated today's grades in physics[b]	5.5	2	3.78 (0.92)	3.78 (0.86)	3.62 (0.85)	3.73 (0.87)
Reported topic knowledge[c]	2	62	19.86 (14.61)	21.79 (12.15)	28.53 (14.91)	23.47 (14.32)
Mean JOL[c]	6	88	56.00 (20.61)	47.79 (14.56)	46.73 (19.34)	50.14 (18.62)
Mean performance[c]	9.5	80.5	48.74 (14.63)	46.43 (13.11)	45.74 (17.40)	46.96 (15.05)
Mean verification performance[c]	10	90	52.59 (15.62)	47.76 (15.50)	48.67 (20.28)	49.66 (17.28)
Mean picture labelling performance[c]	4	72	44.90 (15.67)	45.09 (15.26)	42.82 (17.36)	44.25 (15.99)
Absolute monitoring accuracy[d]	− 47	64	7.26 (16.73)	1.37 (14.14)	0.99 (23.55)	3.18 (18.65)
Relative monitoring accuracy[e]	− 1	1	.42 (.31)	.54 (.37)	.35 (.51)	.44 (.41)
Drawing quality[c]	14.1	86.2	–	44.09 (15.21)	44.32 (17.64)	44.21 (16.35)

[a]possible values ranged from − 10 up to 10

[b]German school grades reach from 1 (very good) to 6 (failed)

[c]Possible values ranged from 0 up to 100

[d]Possible values ranged from − 100 up to 100

[e]Possible values ranged from − s1 up to 1

the scores for monitoring accuracy, of which participants were unaware when making the selection.[1] While working through the post-test, participants neither had access to the paragraphs nor to their drawings from the experimental task. The order of the verification items and picture-labeling tasks was fixed and equivalent to the order in which the paragraphs had been learned before. Finally, participants were debriefed and compensated (with 10 Euro) for their participation. In total, an experimental session took about 75 min.

Data analysis

In line with the typical distinction in metacognition research (de Bruin and Van Gog 2012; Serra and Metcalfe 2009), we relied on two measures of monitoring accuracy to investigate our hypotheses: the absolute monitoring accuracy (calibration) indicating whether the overall level of confidence is adequate, and the relative monitoring accuracy (resolution) indicating whether students could adequately differentiate between which paragraph they learned (not so) well. First, we calculated scores for *absolute monitoring accuracy* by subtracting posttest performance from the JOL magnitudes per participant and paragraph. These scores were then analyzed by means of linear mixed effects models (LMEs) using the R package lme4 (Bates et al. 2015) and lmerTest (Kuznetsova et al. 2017). Our model specification was as follows for absolute monitoring accuracy:

$$Y_{ij} = \alpha_0 + \alpha_1 C_{ij} + \mu_{0j} + \varepsilon_{jj}$$

where Y_{ij} is the JOL minus performance value of person i for text j, α_0 is the grand intercept, $\alpha_1 C_{ij}$ is the regression term of the condition effect (fixed effect), μ_{0j} is the intercept of the text factor (random effect), and ε_{jj} is the error term. We used the Satterthwaite's method for approximating degrees of freedom to obtain t- and p-values for the fixed effects. We used LMEs to analyze absolute monitoring accuracy because they enabled us to analyze the overall condition effect without having to suffer from information loss due to prior averaging over JOLs for the five paragraphs of the instruction (Judd et al. 2012). Rather, using LMEs we could assess the size of the condition effect—which was of substantive interest here (i.e. fixed effect)—accounting for non-independence in the data by simultaneously adding the paragraph number (1–5) to the equation as a random effect (μ_{0j}). Specifying this random effect allowed capturing variances resulting from unspecified differences in JOLs and performance across the texts (e.g., Murayama et al. 2014). Therefore, in this situation LMEs are likely to provide more accurate (and generalizable) estimates of the effects compared to classic ANOVA or regression (e.g., Singmann and Kellen 2017).

Second, as common in studies on comprehension monitoring (Schleinschok et al. 2017; Thiede et al. 2003; van Loon et al. 2014) *relative monitoring accuracy* was operationalized using a Goodman and Kruskal's gamma correlation between the participants' JOLs and their actual performance across texts. It was calculated by first assessing the correspondence between JOLs and performance across the five paragraphs on an *intra*-individual level. For an individual (participant) to reach high accuracy in relative monitoring, the

[1] We refrained from applying inferential statistics to these data, because there was too much data invariance stemming from participants selecting either none or all paragraphs for restudy. This prevented us from calculating intra-individual gamma correlations without having to exclude many participants. Furthermore, the focus of the presented manuscript is on analyzing how monitoring accuracy is affected by our manipulations (in an unconfounded way), which is why there was no actual restudy opportunity for participants.

Table 2 Means (SD) for JOLs and performance scores per paragraph as a function of conditions

	Control	Drawing from scratch	Scaffolded drawing	All subjects
Paragraph 1: JOL	69.66 (18.99)	62.07 (16.56)	54.00 (26.47)	61.82 (21.90)
Paragraph 1: performance	55.17 (18.30)	57.97 (22.71)	56.88 (20.85)	56.68 (20.50)
Paragraph 2: JOL	57.93 (24.26)	50.00 (17.93)	53.00 (22.46)	53.64 (21.72)
Paragraph 2: performance	58.79 (17.20)	52.41 (16.10)	53.75 (23.42)	54.97 (19.21)
Paragraph 3: JOL	46.55 (25.39)	41.03 (18.96)	44.33 (22.54)	43.98 (22.31)
Paragraph 3: performance	27.02 (16.24)	26.19 (16.18)	27.69 (17.41)	26.98 (16.45)
Paragraph 4: JOL	44.83 (25.02)	30.69 (18.70)	32.67 (21.80)	36.02 (22.62)
Paragraph 4: performance	44.40 (21.80)	37.93 (17.99)	38.96 (19.88)	40.41 (19.92)
Paragraph 5: JOL	61.03 (20.76)	55.17 (22.93)	49.67 (21.73)	55.23 (22.08)
Paragraph 5: performance	58.33 (21.83)	57.62 (22.63)	51.44 (27.01)	55.75 (23.90)

paragraphs for which relatively higher JOLs were provided should go along with relatively higher performance in the post-test questions about this paragraph and vice versa, which is indicated by a positive value of gamma (maximum = + 1). If paragraphs with relatively high JOLs go along with relatively low performance, negative values of gamma (minimum = − 1) are the result. To calculate gamma, both the five JOLs and the five performance scores per paragraph (within each participant's data) were ordered in two ordinal scales with values from 1 to 5 to compare the similarity of the orderings when ranked by quantities. These two ordinal scales were correlated nonparametrically (using gamma correlation), yielding a value between − 1 and + 1 per participant. These values were then compared *between* participants from the different experimental conditions. The advantage is that these values are not affected by overall test performance or individual tendencies regarding confidence (Thiede et al. 2003), which is why they are considered an appropriate to measure relative accuracy (Nelson 1984).

Results

Data were analyzed and will be reported in a three-step manner. First, we tested for whether the participants from the three conditions were similar in their entry characteristics. Second, we checked for whether the drawing manipulation was successful. Third, we tested the monitoring accuracy hypothesis followed by the use-of-diagnostic-cue hypothesis. A significance level of 0.05 was used for all analyses. We used Cohen's *d,* Pearson's *r* and ω^2 as effect-size measures (Cohen 2013). For Cohen's *d*, values of 0.20, 0.50, and 0.80 are considered small, medium, and large effect sizes, respectively. For Pearson's *r*, values of 0.10, 0.30, and 0.50 are considered small, medium, and large effect sizes, respectively. For ω^2, values of 0.01, 0.06, and 0.14 are considered small, medium, and large effect sizes, respectively. Tables 1 and 2 show all descriptive values.

Entry characteristics

Analyses of variance (ANOVAs) revealed no significant differences between the conditions concerning participants' age, visuo-spatial abilities, latest school grades in physics, and self-reported physics knowledge (all $F_s < 1$, see Table 1 for descriptive data). There

was a significant difference between conditions regarding the participants' self-reported topic knowledge, $F(2, 85) = 3.16$, $p = 0.047$, $\omega^2 = 0.047$. The comparison of descriptive data revealed that participants in the scaffolded-drawing condition estimated their knowledge as being somewhat higher than those in the drawing-from-scratch and in the control condition but this difference failed to reach statistical significance in a post-hoc comparison test with Bonferroni adjustment ($p = 0.056$). Also, participants' self-reported topic knowledge was not significantly related to any of the dependent variables JOLs, performance, drawing quality, absolute and relative monitoring accuracy (all $p_s > 0.05$). Therefore, we refrained from including it as covariate in the subsequent analyses.

Manipulation check

To consider the drawing manipulation as successful for the present research purposes, two conditions ought to be met. First, there should be sufficient variance in the quality of students' drawings (no floor or ceiling effects) to allow differentiation. Scores for drawing quality ranged between 14.1 and 86.2 ($M = 44.2$, $SD = 16.4$) on a scale from 0 to 100. Hence, students produced drawings of medium quality. There were no floor or ceiling effects.

Second, the drawing scores should be diagnostic for performance, meaning they are correlated with the posttest scores. The drawings turned out to be diagnostic cues for later test performance ($r = 0.43$, $p < 0.001$). Interestingly, when considered separately in the two drawing conditions, the quality significantly related to performance only when drawings were made with scaffolds ($r = 0.70$, $p < 0.001$), but not when made without scaffolds ($r = 0.18$, $p = 0.35$). The difference between both correlation coefficients was significant, $z = -2.49$, $p = 0.006$.

Monitoring accuracy hypothesis

To test the hypothesis as to whether (scaffolded) drawing would support *absolute* monitoring accuracy, we calculated a linear mixed effects model in which JOLs minus performance for each of the five paragraphs were entered as the dependent variable (see Table 2 for descriptive data), the experimental condition (control vs. drawing vs. scaffolded drawing) as fixed effect, and the paragraph number (1 to 5) as random effect. Within this model, we translated our two hypotheses into a Helmert contrast, meaning that we first expected the control condition to show higher (overconfidence) values than both of the drawing conditions (contrast code: $+2; -1; -1$). Results revealed significant differences in line with this contrast, $b = 2.84$, $SE = 0.81$, $t = 3.50$, $p < 0.001$, $d = 0.36$ (see Fig. 3). Second, we expected the drawing (without scaffold) condition to have higher values than the scaffolded drawing condition ($0; +1; -1$). Results were not in line with this assumption, as they revealed no significant difference between the two drawing conditions on absolute monitoring accuracy, $b = 1.27$, $SE = 1.39$, $t = 0.91$, $p = 0.36$, $d = 0.10$. The paragraph number, specified as random effect within the model, explained 8.2% of the total variance. This can be considered a medium effect (cf. LeBreton and Senter 2008) that warrants separate inclusion in the model.

Concerning *relative* monitoring accuracy, we analyzed the values for (intra-individually obtained) gamma correlations and also tested our hypothesis by means of whether these values differ between conditions according to Helmert contrasts ($[-2; +1; +1]$, $[0; -1; +1]$). Results revealed that gamma values did neither differ significantly

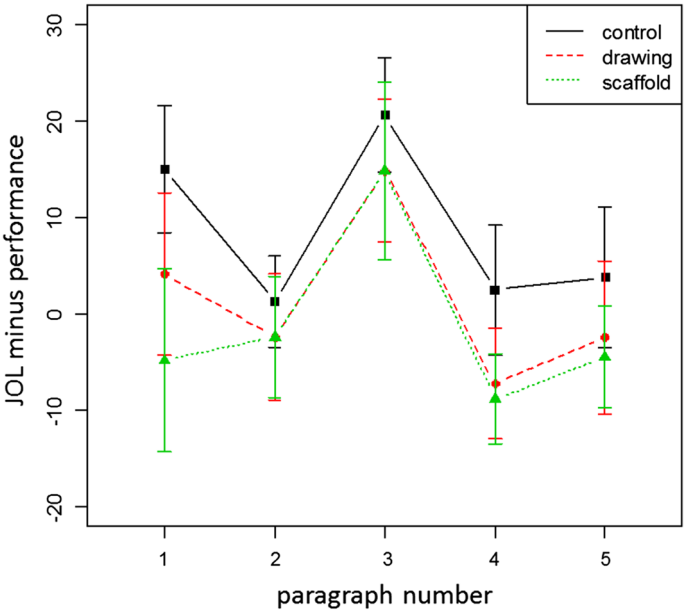


Fig. 3 Scores and standard errors for absolute monitoring accuracy (JOL-performance) across the five paragraphs of the text, and as a function of experimental condition: control (solid black line) vs. drawing-from-scratch (dashed red line) vs. scaffolded-drawing (dotted green line). A value of 0 indicates perfect accuracy. (Color figure online)

between the control and the two drawing conditions, $b = -0.09$, $SE = 0.08$, $p = 0.28$, $d = -0.26$, nor within the two drawing conditions, $b = 0.19$, $SE = 0.10$, $p = 0.052$, $d = 0.49$—although in the latter case, the comparison just missed significance. Cautiously interpreting this finding, contrary to our expectation gamma values were (descriptively) higher for participants in the drawing-from-scratch than in the scaffolded-drawing condition.

Use-of-diagnostic-cue hypothesis

We conducted a moderation analysis with mean z-standardized drawing qualities across paragraphs, the two drawing conditions (drawing-from-scratch vs. scaffolded-drawing), and their interaction term as predictors and mean JOLs across paragraphs as dependent variable. This was done to test whether students actually based their JOLs more strongly on whether they drew the crucial elements in the scaffolded-drawing than in the drawing-from-scratch condition. Results revealed that the expected interaction missed significance, $b = 2.81$, $SE = 2.20$, $t = 1.28$, $p = 0.21$, $r = 0.17$. A regression analysis without the interaction term, however, revealed that drawing quality was a significant predictor for JOLs, $b = 0.29$, $SE = 0.13$, $t = 2.22$, $p = 0.03$, $r = 0.28$, whereas the condition was not, $b = -1.13$, $SE = 4.33$, $t = -0.26$, $p = 0.80$, $r = 0.03$. Thus, drawing quality was used as cue for metacognitive monitoring to a similar extent in both drawing conditions.

Discussion

The present study sought to replicate and expand findings from previous research on whether drawing supports metacognitive monitoring. Specifically, we tested whether monitoring accuracy can be supported even more strongly when students are scaffolded in their drawing efforts prior to providing their monitoring judgments (JOLs). To this end, we compared results for monitoring accuracy between a control condition, a drawing-from-scratch condition and a scaffolded-drawing condition.

As in previous research (e.g., Schleinschok et al. 2017), we expected that the drawing tasks (with and without scaffold) would help students base their JOLs on cues that are diagnostic for performance, resulting in more accurate metacognitive monitoring. We found partial support for this Monitoring Accuracy Hypothesis. Although there were no differences between the conditions regarding relative monitoring accuracy, data shows the expected pattern regarding absolute monitoring accuracy. The linear mixed effects model revealed significant differences between the control condition and the two drawing conditions. Students who drew after learning (both from scratch and scaffolded) showed better absolute monitoring accuracy (calibration; cf. Alexander 2013) than students who should, among others, think about which keywords describe the previously read text in the control condition (cf. Fig. 3). The latter is a rather strong control condition as generating keywords have been shown to improve the (relative) accuracy of judgments (Thiede et al. 2003). At first sight, this finding corroborates previous research in that drawing supported monitoring (Schleinschok et al. 2017); moreover, it extends the previous research by revealing the drawing benefits compared to a control condition, in which time-on-task was identical to the drawing conditions. However, Schleinschok et al. (2017) found benefits of drawing from scratch on *relative* monitoring accuracy whereas both drawing from scratch and scaffolded drawing were beneficial only to *absolute* monitoring in the present study. Moreover, unlike expected there were no differences between the two drawing conditions (from-scratch vs. scaffolded) on relative or absolute monitoring accuracy. A plausible reason for the latter non-significant finding is that drawing from scratch already led to near perfect accuracy (mean scores close to 0; see Fig. 3 and Table 2) so that there was not much room for further improvement in the scaffolded-drawing condition.

Reasons for why no benefits of (scaffolded) drawing on relative monitoring accuracy were found may lie in the fact that there were medium to large differences in difficulty between the text sections. As can be seen in Fig. 3, students from all conditions overestimated their performance for text Sect. 3, because performance was lowest for this section. It was obviously more difficult than the others. Such differences in difficulty between the sections might have contributed more strongly to scores for relative monitoring accuracy (i.e., being able to differentiate which sections are better or less well understood) than the drawing manipulations. The overall score of $M = 0.44$ ($SD = 0.41$) for gamma correlations in this study tentatively supports this assumption. Provided that, unlike in much of the previous research (e.g., Thiede et al. 2003; van Loon et al. 2014; Pieger et al. 2016), our gamma correlations were based on JOLs from different sections of the same text (not from multiple unrelated texts)—with varying difficulties—the scores for gamma correlations are quite high. This might have occluded potential effects of the drawing manipulation on relative monitoring accuracy. Hence, we recommend further research on this topic using texts or text sections of similar difficulty to potentially find effects of (scaffolded) drawing on relative monitoring accuracy.

Another aspect to be considered when trying to explain the non-significant differences between drawing and scaffolded drawing on (absolute and relative) monitoring accuracy is a potential assistance dilemma (Koedinger and Aleven 2007). Our hypothesis was based on the premise that there was an optimal level of assistance in the scaffolded-drawing condition, whereas there was too little assistance in the drawing-from-scratch condition. Conversely, one might conclude from the present results that there was too much assistance in the scaffolded-drawing condition so that monitoring accuracy was not further supported. Accordingly, if there was even more assistance, for instance in the form of provided drawings or diagrams, previous research has revealed that this could inflate rather than sharpen metacognitive judgments (Eitel 2016; Serra and Dunlosky 2010). This was not the case here, as the scaffolds consisted only of outlines providing a frame onto which subsequent information should be added. Nevertheless, the scaffolds that were used here might not have provided an optimal level, but too much support; hence, further research is welcomed that uses varying degrees of support by spatial scaffolds, and also assesses online self-monitoring behavior to better understand the processes behind the effects of (scaffolded) drawing (cf. Van Meter 2001).

According to our Use-of-Diagnostic-Cues Hypothesis, students were expected to base their JOLs on whether they were able to draw a coherent image of the crucial elements especially after drawing with a spatial scaffold. Results partially confirm this hypothesis as the drawing quality significantly predicted JOL magnitudes in both drawings conditions taken together (when no interaction was modelled). Students in both drawing conditions similarly based their metacognitive monitoring, among others, on the actual quality of their drawings. This might explain why there were comparable beneficial effects on metacognitive monitoring in both drawing conditions compared to the control condition without a drawing task. Drawing with and without scaffold seemed to have acted as feedback for how well one has understood the text contents in the present research. Nonetheless, the present results did *not* reveal that students based their JOLs more strongly on the quality of drawings made with spatial scaffolds compared to drawing from scratch. Unlike in previous research on open-book drawing (cf. Leutner et al. 2009; Schmidgall 2017), the scaffolds did not seem to foster more focused drawing by prompting where to draw and (in part) what to draw. As mentioned above, one potential reason is that an optimal level of assistance was exceeded by the scaffolds.

Limitations and conclusions

We used instructional materials that required a high degree of visuo-spatial reasoning to be understood (cf. Fig. 1). It is still an open question whether a visuo-spatial task such as drawing can support monitoring accuracy only for visuo-spatial contents as used here. In other words, it remains to be seen whether the present findings generalize to contents lower in visuo-spatial reasoning demands in future studies. One might expect that drawing as a visuo-spatial task fosters monitoring especially for high visuo-spatial contents whereas keyword or summary writing (verbal tasks) might work better for low visuo-spatial contents. We thus invite further research investigating how the type of generative task interacts with the type of study contents to affect metacognition and learning.

As a limitation it is to note that we used verification and picture labeling tasks but no open questions to assess performance. This was done to keep writing demands for participants on a reasonable level reducing error variance due to differences in students'

motivation to write, and to allow for highly objective data scoring. While it is generally possible to assess student's comprehension by having them work with closed item formats such as multiple-choice or verification (Lindner et al. 2015; Simkin and Kuechler 2005), the typical format to assess deeper comprehension and transfer that reflect situation-model reasoning is an open one. In the present study, we sought to assess comprehension by formulating closed items in a way that correct responses required an inference to be drawn. Hence, we welcome future research administering constructed-response and essay tasks.

Future research may also tackle whether differences in monitoring accuracy can be explained by differences in test expectancy that were stimulated by the different tasks (cf. Thiede et al. 2011). The drawing task might have induced a higher expectancy for tasks asking for the relations between major concepts, and therefore tasks requiring situation-model reasoning more strongly than the control condition. However, this is so far speculative and needs further research. In such research, it would be helpful to include an additional control group to the design where students are specifically instructed to mentally reactivate the text contents while focusing on important relationships between major concepts—similar to what students in the drawing conditions might have focused on.

Notwithstanding the above, the present results suggest that drawing, both with and without spatial scaffolds, can support metacognitive monitoring for high visuo-spatial learning contents. Moreover, spatial scaffolds seemed to help students base their metacognitive judgments more strongly on the quality of their drawings, and thus on a cue that is diagnostic for learning and performance.

Appendix

Paragraph 2: the Earth's magnetic field

The second phenomenon that plays an important role in the formation of auroras is the earth's magnetic field, that is, the magnetic field that surrounds the earth. The earth's magnetic field has two poles, the magnetic north and the magnetic south pole. This magnetic poles do not exactly coincide with the geographic poles. The magnetic poles describe the places where the earth's magnetic field enters the earth vertically. The geographic poles, however, describe the places where the imaginary earth's axis (axis of rotation) penetrates the surface of the earth. The earth's rotation axis is currently inclined around 11.5 degrees from the axis of the earth's magnetic field. The magnetic field lines of the earth's magnetic field run from the southern hemisphere to the northern hemisphere. Without the solar wind, the magnetic field lines of the earth's magnetic field would encircle the earth in controlled arches (comparable to a bar magnet). However, the earth's magnetic field is deformed by the solar wind, which is described in more detail later. The area around the earth that is enclosed by the earth's magnetic field is called the magnetosphere. The third phenomenon that plays a role in the formation of auroras are specific properties of magnetic fields. One such property is that electrically charged particles cannot move crosswise through magnetic fields. If an electrically charged particle hits a magnetic field line, it is forced into a spiral path around the magnetic field lines. Thus, it cannot cross the magnetic field and instead moves along the magnetic field lines around the magnetic field. In space, this is an important principle since there are almost only charged particles and (albeit weak) magnetic fields.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The authors confirm, that all measures, conditions, and data exclusions of the study were reported in this paper.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures followed were in accordance with the ethical guidelines of the German Psychological Society's (DGPs; 2004, CIII) and APA standards. All participants provided informed consent prior to taking part in this study. They were aware of taking part in research and informed about the possibility of quitting the experiment with no repercussions out disadvantage at any time. All participants allowed us to use their data for publications. All data were collected and analyzed anonymously. The study was part of a larger project about the effects of drawing on self-regulated learning that received ethical approval by the Internal Review Board of the Leibniz-Institut fuer Wissensmedien.

References

Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction, 24*, 1–3. https://doi.org/10.1016/j.learninstruc.2012.10.003.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v067.i01.

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.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hoboken: Taylor and Francis. Retrieved from https://gbv.eblib.com/patron/FullRecord.aspx?p=1192162.

de Bruin, A. B. H., & van Gog, T. (2012). Improving self-monitoring and self-regulation. From cognitive psychology to the classroom. *Learning and Instruction, 22*, 245–252. https://doi.org/10.1016/j.learninstruc.2012.01.003.

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.

Eitel, A. (2016). How repeated studying and testing affects multimedia learning: Evidence for adaptation to task demands. *Learning and Instruction, 41*, 70–84. https://doi.org/10.1016/j.learninstruc.2015.10.003.

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). *Manual for kit of factor referenced cognitive tests*. Princeton: Educational Testing Service.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146.

Hegarty, M., & Waller, D. (2005). Individual differences in spatial abilities. In P. Shah & A. Miyake (Eds.), *The Cambridge handbook of visuospatial thinking* (pp. 121–169). Cambridge: Cambridge University Press.

Hilbert, T. S., & Renkl, A. (2008). Concept mapping as a follow-up strategy to learning from texts: What characterizes good and poor mappers? *Instructional Science, 36*(1), 53–73. https://doi.org/10.1007/s11251-007-9022-9.

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.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*(1), 54–69. https://doi.org/10.1037/a0028347.

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.

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review, 19*(3), 239–264. https://doi.org/10.1007/s10648-007-9049-0.

Koriat, A., & Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition, 34*(5), 959–972. https://doi.org/10.3758/BF03193244.

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

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tessts in linear mixed effects models. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v082.i13.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815–852. https://doi.org/10.1177/1094428106296642.

Lenhard, W., & Lenhard, A. (2014). *Berechnung des Lesbarkeitsindex LIX nach Björnson*. Verfügbar unter: https://www.psychometrica.de/lix.html. Bibergau: Psychometrica

Leutner, D., Leopold, C., & Sumfleth, E. (2009). Cognitive load and science text comprehension: Effects of drawing and mentally imagining text content. *Computers in Human Behavior, 25*(2), 284–289. https://doi.org/10.1016/j.chb.2008.12.010.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26*(2), 149–171. https://doi.org/10.1016/0030-5073(80)90052-5.

Lindner, M. A., Strobel, B., & Köller, O. (2015). *Multiple-Choice-Prüfungen an Hochschulen? Ein Literaturüberblick und Plädoyer für mehrpraxisorientierte Forschung [Are multiple-choice exams useful for universities? A literature review and argument for a more practice oriented research]*. Zeitschrift für Pädagogische Psychologie, *29*, 133–149. https://doi.org/10.1024/1010-0652/a000156.

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 40*(5), 1287–1306. https://doi.org/10.1037/a0036914.

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. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–173). New York: Academic Press.

Pieger, E., Mengelkamp, C., & Bannert, M. (2016). Metacognitive judgments and disfluency: Does disfluency lead to more accurate judgments, better control, and better performance? *Learning and Instruction, 44*, 31–40. https://doi.org/10.1016/j.learninstruc.2016.01.012.

Rawson, K. A., & Dunlosky, J. (2002). Are performance predictions for text based on ease of processing? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 69–80. https://doi.org/10.1037/0278-7393.28.1.69.

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.

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cognitive Science, 38*(1), 1–37. https://doi.org/10.1111/cogs.12086.

Scheiter, K., Schleinschok, K., & Ainsworth, S. (2017). Why sketching may aid learning from science texts: Contrasting sketching with written explanations. *Topics in Cognitive Science, 9*, 866–882. https://doi.org/10.1111/tops.12261.

Schleinschok, K., Eitel, A., & Scheiter, K. (2017). Do drawing tasks improve monitoring and control during learning from text? *Learning and Instruction, 51*, 10–25. https://doi.org/10.1016/j.learninstruc.2017.02.002.

Schmidgall, S. (2017). *Drawing to learn: Investigating the role of contributing factors and instructional support for learner-generated drawing* (Doctoral dissertation, Eberhard Karls Universität Tübingen).

Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgements reflect the belief that diagrams improve learning from text. *Memory, 18*, 698–711. https://doi.org/10.1080/09658211.2010.506441.

Serra, M. J., & Metcalfe, J. (2009). Effective implementation of metacognition. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 278–298). New York: Routledge.

Simkin, M. G. & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education*, *3*, 73–97. https://doi.org/10.1111/j.1540-4609.2005.00053.x.

Singmann, H., & Kellen, D. (2017). An introduction to mixed models for experimental psychology. In D. H. Spieler & E. Schumacher (Eds.), *New methods in neuroscience and cognitive psychology*. Hove: Psychology Press.

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

Thiede, K. 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., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology, 81*(2), 264–273. https://doi.org/10.1348/135910710X510494.

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

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.

Van Meter, P. (2001). Drawing construction as a strategy for learning from text. *Journal of Educational Psychology, 93*(1), 129–140. https://doi.org/10.1037/0022-0663.93.1.129.