Article

Effects of Strategy Training and Elaboration vs. Retrieval Settings on Learning of Cell Biology Using Concept Mapping

Lukas Bernhard Becker [†] [ID], Virginia Deborah Elaine Welter [†] [ID] and Jörg Großschedl *[ID]

Faculty of Mathematics and Natural Sciences, Institute for Biology Education, University of Cologne, D-50931 Cologne, Germany; l.becker@uni-koeln.de (L.B.B.); virginia.welter@uni-koeln.de (V.D.E.W.)
* Correspondence: j.grossschedl@uni-koeln.de
† These authors contributed equally to this article.

Abstract: Concept mapping (CM) can be an effective strategy to facilitate understanding-based learning of complex and abstract scientific content. In our study, we tried to answer two questions that have not yet been clarified: (1) How effective is CM if the learners do not have the textual learning material available when constructing their concept maps (retrieval setting) in contrast to when the material is available (elaboration setting)? (2) Do potential effects of the two settings depend on the intensity of a previous CM strategy training? To this end, we established a quasi-experimental four-group plan and collected data of $N = 93$ undergraduate students. The participants received either a detailed CM strategy training or just a short introduction and were then asked to apply CM in a retrieval or elaboration setting. The quality of constructed concept maps and content-related learning success were assessed to determine potential group differences. Overall, our results show a significantly positive but setting-independent effect of the CM strategy training on methodical CM skills. However, the different setting conditions differed neither regarding concept map quality nor content-related learning success, implying that CM in retrieval settings could be as effective as in elaboration settings.

Keywords: concept mapping; strategy training; retrieval practice; elaboration; organization; knowledge acquisition; cell biology

1. Introduction

As a result of learners' trouble with learning in STEM subjects such as biology [1–4], researchers and educators have tried to identify new methods of instruction to afford learners the best possible learning success [5–11]. While it can certainly be helpful for learners to visualize abstract systems such as the inside of a cell with an animation or even with an augmented reality device, the question remains how learners process information, how they organize it, and how they can learn in an active and meaningful way in contrast to mere passive reception [12]. Answering these questions is crucial since the learners' challenge in biology, as in all STEM subjects, is not to merely memorize individual chunks of information but to think in extensive contexts by connecting these chunks, and thus, gain a deep understanding of their overall interrelationships [7,13–16]. This, in turn, can help them to detect interdependencies between different levels of conceptual organization, for example, regarding different biological systems. In biology, such interdependencies are also typical for the area of cell biology, which is characterized, among other things, by complex interconnected concepts ranging on different system levels [17]. However, the resulting complexity of information can pose a challenge for learners, as the cognitive processes of organization and elaboration play a crucial role for understanding such relationships [18–24].

In the context of text-based learning, cognitive processes of organization involve recognizing which statements of a section and which connections between individual chunks of information/hierarchical elements are of particular importance. Organization processes

can be stimulated by appropriate cognitive strategies, including underlining key messages in a text, writing summaries, or creating illustrations [25]. By applying such strategies, learners should become aware of existing connections between individual chunks of information. Such systematic organization of knowledge can also support its reconstruction when reproduction is required (e.g., within an exam). In contrast, cognitive processes of elaboration refer to joining new information and prior knowledge elements [26,27]. Corresponding elaboration strategies therefore include verbal or pictorial enrichment, linking to examples from everyday life and personal experience, or building analogies. Using such strategies can stimulate elaboration processes, which are essential for meaningful and understanding-based learning [28,29] and facilitate connected thinking in the context of science education [30].

Complementing the research on learning efficacy of cognitive strategies such as organization and elaboration, increasing attention has also been paid to the mechanism of retrieval of information from memory and specifically to a possible retroactive effect of this retrieval on memory organization and the learning performance [31–35]. Retrieval processes are assumed to occur when learners retrieve information from memory without simultaneous availability of the learning material, resulting in a setting of retrieval practice [36–39]. In this regard, studies could show that such retrieval practice not only provides learners with an opportunity to monitor their learning progress but improves learning itself, as well [39–42]. Compared to retrieval, the operationalization of elaboration processes is usually characterized by a simultaneous availability of learning material, which the learners actively deal with in order to integrate new information into their prior knowledge. Following this strain of reasoning, such a setting is called elaborative studying setting.

Consequently, a plausible approach to promote learning success could lie in using an organization or elaboration strategy within a retrieval setting. However, according to O'Day and Karpicke [43], studies combining retrieval practice and elaborative learning strategies are "woefully sparse" (p. 2).

1.1. Concept Mapping

One learning strategy that has been used for both organization and elaboration but has rarely been used specifically in a retrieval setting is concept mapping (CM). Previous findings indicate that CM facilitates learning of abstract topics and complex relationships in STEM subjects, such as biology [44–48], physics [49,50], and chemistry [51,52]. Products of applying CM, the concept maps, are network diagrams, representing the types of relationships between meaningful terms or concepts. In practical terms, such a network is created by connecting the concepts (nodes) with labelled arrows: while the label indicates the semantic relation between two concepts, the arrow direction indicates the reading direction. Two concepts connected by a labelled arrow constitute a so-called proposition, representing the smallest meaningful unit of a concept map. Areas of concept maps that are not in close proximity can be associated with each other via building cross-connections to point out an existing relationship [53–58].

The aforementioned features suggest concept maps to be an isomorphic analogy to the assumed structure of memory, which is based on Quillian's semantic network model [59,60]. According to this model, knowledge is represented in more or less hierarchically structured semantic networks of memory [61,62]. Relationships connecting individual concepts that form the structure of these memory networks are semantic, logical, and grammatical in nature. As concept maps simplify the cognitive processing of semantic relations by explicating the logical elements between one concept and another, CM is particularly suitable to facilitate an understanding of complex issues. The specification of complex relationships during CM requires a systematic and analytical approach, leading to structured representations of contexts in resulting concept maps [58,63–65]. In addition, these processes during CM can promote organization and elaboration processes [66] and encourage learners to analyze the learning material on a deeper level, resulting in increasing learning success [44,53,67–73].

Furthermore, CM can also serve as a metacognitive tool, as noticing one's own understanding and misunderstanding is particularly important for developing metacognitive skills, which has been named one of the main arguments in favor of teaching students CM [74–78]. By drawing the mapper's attention to difficulties regarding the plausible integration of concepts or the specification of relationships, concept maps can help learners to recognize knowledge gaps or flawed logic, so they are more likely to become able to react to them, for example, by restudying the material.

Finally, CM can be used to assess learning success, e.g., by analyzing the number of propositions specified by learners. From a pedagogical point of view, concept maps can serve as such a diagnostic tool because they represent the learner's individual understanding of a respective domain [79,80]. Therefore, a concept map draws the teacher's attention to what the learner (already) knows, but also to what the learner did not understand or what may have been misunderstood. In this regard, research has shown that concept maps are useful in assessing how learners relate, organize, and structure concepts [81]. Therefore, it is not surprising that CM has been used as a diagnostic tool in many studies regarding meaningful learning [82–86].

1.2. Previous Research on the Effectiveness of Concept Mapping in Retrieval Settings

Since CM stimulates the aforementioned processes of organization and elaboration in the course of working on learning material [87], the experimental design of constructing concept maps with simultaneous availability of a learning text (elaboration setting) seems the obvious configuration to enable the learners to understand the text using CM before their learning success is measured [38,71,88].

An alternative design was used by Blunt and Karpicke [37] to explore the effectiveness of CM, a typical elaborative learning strategy, in a retrieval practice setting. In two experiments, they were able to show that CM can also be used effectively within retrieval practice formats and that the two established learning formats, CM vs. note-taking, each resulted in better learning outcome in terms of factual and inferential knowledge when used in a retrieval practice setting instead of a classical elaboration setting in which the learning material is available. Blunt and Karpicke primarily attribute this learning strategy-independent finding to generally better skills of the retrieval practice groups' participants regarding the use of retrieval cues for knowledge reconstruction [37].

For further exploration of the learning effectiveness of CM as an elaboration-supportive learning strategy in a retrieval setting, O'Day and Karpicke [43] conducted another two experiments in which students had to read short texts and practiced retrieving the information by free recall, CM, or both. Their results indicate a superiority of the free recall group over the CM group and the CM-and-recall group regarding retention of information.

However, the aforementioned results of Blunt and Karpicke [37] as well as O'Day and Karpicke [43] should be viewed with caution since they could have been biased by a lack of practice and familiarity with CM on the part of the learners, causing additional cognitive load. These studies' participants who were asked to apply CM in learning only received a short introduction about the strategy's basics, which has also been criticized by researchers such as Mintzes et al. [89]. In addition, the brevity of the learning text of approximately 250 words used in the study of O'Day and Karpicke [43] may have represented an unrealistically low burden for students and could therefore have undermined their motivation to elaborate on their prior knowledge.

Since CM is also considered an effective metacognitive tool (see Section 1.1), Karpicke and Blunt [38], Blunt and Karpicke [37], and O'Day and Karpicke [43] additionally examined the aspect of judgement of learning (JOL) by asking their participants to judge how much (0 to 100%) of the learning content they will be able to remember one week later. The results of Karpicke and Blunt [38] and Blunt and Karpicke [37] show that even though retrieval practice groups performed better in follow-up tests, students in the elaborative studying setting rendered higher JOLs beforehand. Comparable results were also found in other studies focusing on effects of the availability of learning material [90,91]. The

incorrect prognoses regarding one's own learning success are often explained by the empirical finding that JOLs apparently show a positive linear dependency on the perceived ease of information processing. When the material is available to learners, information processing usually seems smooth and easy, resulting in high JOLs and somewhat unrealistic self-efficacy beliefs compared to actual test performance levels. In contrast, it is possible that active retrieval changes the basis of judgement: instead of judging how easily information can be read and/or processed, judgments could more likely be based upon the ease of memory recall of this information during retrieval practice [92–94].

1.3. Previous Research on the Effectiveness of Different Concept Mapping Training Approaches

Since CM requires abilities such as identifying important terms or concepts (e.g., in a learning text), determining hierarchical relationships among them, and specifying meaningful propositions, it seems obvious that learners need additional training and support to use the strategy successfully. Accordingly, some studies involved several weeks of training with repeated intervention and feedback measures [44,45,78,83,89,95,96]. In contrast, other researchers such as Jonassen et al. [97] state that CM is comparably easy to learn, so they regard short introductions to be sufficient when investigating the strategy [37,38,43]. Even if such short introductions have been implemented successfully in some cases [98], it is not clear to what extent such effects depend on specific characteristics of the participants (e.g., cognitive abilities, previous experience with CM) and/or methodical variations between the studies (e.g., level of requirement, intensity of instruction). Accordingly, empirical recommendations on parameters such as the content and duration of a training course/instruction vary [84,86,99,100].

In a previous study, we already have been able to show that the overall learning success can be improved by an extensive CM training and that stable CM strategy skills, in particular, can be promoted by additional integration of scaffolding and feedback elements in the course of such a training. However, we also noted that our small sample (N = 73) was not sufficient in terms of external validity and further studies were needed to replicate these findings [44]. Therefore, in the present study, we wanted to take the opportunity to validate our previous results.

1.4. Research Questions and Hypotheses

Considering the aforementioned different training and instruction approaches as well as the different effects of CM setting variations (elaboration vs. retrieval) on learning outcome, we designed a study to provide more clarity in this regard. Specifically, we try to answer the following questions, using a four-group plan: (1) How effective is CM if the learners do not have the textual learning material available while constructing their concept maps (retrieval setting) in contrast to the classical elaboration setting, where the learning material is available? (2) Do potential effects of the two settings depend on the intensity of a previous method training in the utilization of CM? To answer these two questions, we have translated them into three corresponding hypotheses:

(1) Based on the findings of Blunt and Karpicke [37], we expect that the availability of the learning material is crucial for cognitive processes during CM. In absence of a learning text, a CM construction task regarding the learning content should induce memory-related recall processes which bind cognitive resources usually required for further elaboration. Vice versa, elaboration processes should increase if the learning material is available to participants during CM, which should also be reflected in a more elaborative character of the propositions participants specified in their concept maps. Since a higher level of elaboration is additionally associated with better learning outcomes in general, we expect these participants to achieve higher scores in subsequent knowledge tests regarding the covered learning content.

(2) Furthermore, we expect that extensive training in CM is helpful for its successful use as a learning tool [44,45,95] by inducing a certain familiarity, especially with methodical aspects. This effect should become particularly evident in comparison with

another condition in which the participants only receive a short introduction to the CM strategy. Accordingly, we expect participants who took part in an extensive CM training to show higher CM-related self-efficacy expectations, to achieve better results in analyzing and editing given concept maps, as well as to make fewer methodical mistakes when creating own concept maps.

(3)   Finally, we expect that the combination of both factors, an extensive CM training and the availability of the learning material during CM, will improve CM- and knowledge-related learning outcomes the most, which should also be reflected in a related metacognitive assessment of the participants regarding their own learning success [36].

2. Materials and Methods

Our quasi-experimental study was conducted as a four-group plan. On the one hand, the variable "CM training intensity" (extensive CM training [T+] vs. control training followed by a short introduction to CM [T−]) was varied; on the other hand, the variable "availability of learning material" during CM (learning material not available/retrieval [R] vs. learning material available/elaboration [E]) was varied. The implementation of this design took place over a six-week period of weekly sessions, which comprised three consecutive phases:

(1)   A training phase (weeks 1–3), in which students received an extensive CM training (T+) or a control training (T−);
(2)   A learning phase (week 4), in which students created a concept map on the topic cell biology in presence (E) or absence (R) of the respective learning material; and
(3)   A test phase (weeks 5–6), in which the students' learning success was measured by implementation of various knowledge tests (see Section 2.4).

In order to avoid test effects of the training phase on the later learning and test phases, we have chosen different but comparably abstract learning topics for these parts: intelligence for the training phase and cell biology for the learning and test phases. This approach has proven to be useful in our previous study [44].

The learning texts that the participants worked on during the study units of the training and learning phases were designed by us prior to our previous study [44] and assessed as appropriate to the university level by 13 experts in the respective fields of biology education ($n = 9$) and psychology ($n = 4$). The text Theories and Models of Intelligence comprised 3197 words (8 pages) and was divided into three equal units to cover the three training sessions, whereas the text The Structure and Function of Eukaryotic Cells comprised 2010 words (7 pages) and was used entirely in the learning phase. Throughout the study, the same standard of behavior and situation was strictly implemented for all four groups in every study-related interaction with the participants in order to minimize investigator and context effects.

2.1. Sample

A total of $N = 93$ undergraduate university students of different fields of study (54% enrolled in a natural science study program) participated in our study. On average, our participants were 21.9 years old, and 84% of them were female.

Participation was based on self-selection in the first instance, as our study was carried out as part of an elective curricular course on learning strategies. Accordingly, the participants had to spend their free time, so we offered each weekly session at a total of five different time slots (different days of the week and different times of the day). After all participants had individually decided in advance on the most suitable time slot for them, they were permanently assigned to it for the entire duration of the study. Afterwards, we randomly assigned each of the five time slot groups to one of the two training conditions, resulting in a total of $n = 48$ participants in the T+ and $n = 45$ participants in the T− groups. These participants were again randomly assigned to one of the two setting conditions (learning material not available/retrieval [R] vs. learning material available/elaboration

[E]), so that the four final quasi-experimental groups were created (see Figure 1). The minimally unequal distribution of the participants in the respective groups is due to a dropout of a total of five subjects during the study period.
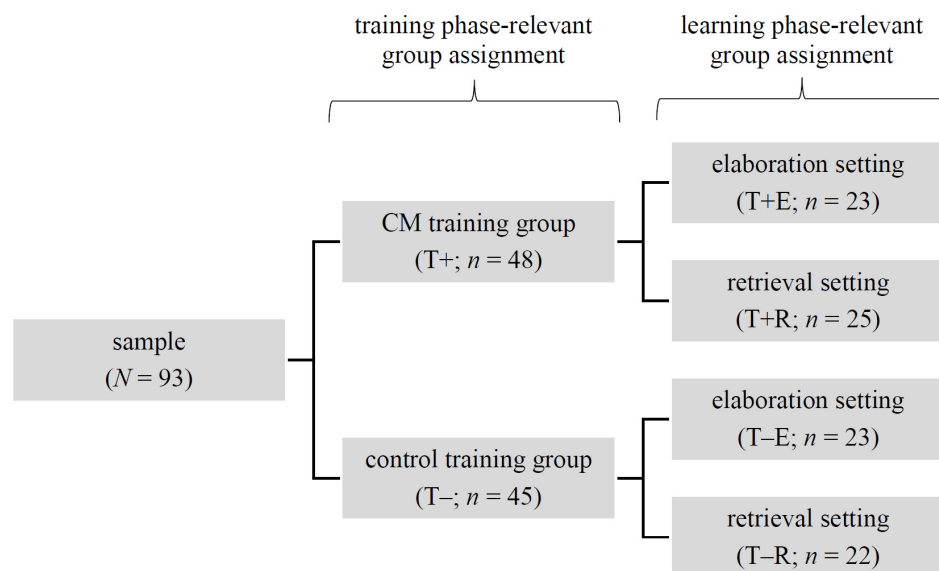

Figure 1. Composition of study groups.

2.2. Procedure

During the three training sessions (weeks 1 to 3), all participants worked on a learning text on the topic of intelligence, but the groups differed in terms of the strategy they should use in learning: whereas the T+ groups used CM, the T− groups used non-CM strategies. Accordingly, the participants in both groups received different instructions, which were partly based on that of our previous study, and thus, already proven to be useful [44]:

(1) Participants of the T+ groups received an extensive CM training including supportive and feedback elements. Every weekly session started with a theoretical lecture on the CM strategy. In addition to the lectures, the participants received CM-specific scaffolding (see Table 1) and metacognitive prompts regarding the individual study unit following the lecture (e.g., "Did I label all arrows clearly, concisely, and correctly?"; adapted from Großschedl and Harms [101,102]). However, the number of prompts was reduced over the course of the training phase (fading) [73] to prevent unnecessary distraction. In order to give the participants the opportunity to check the correctness of the concept maps they had constructed, they received a scripted overview of the most common CM errors. Feedback by the instructor was provided continuously during the study unit: individual verbal feedback was given on request during the participants' construction of own concept maps, and written feedback on these constructed concept maps was given after the study unit using a knowledge of correct results (KCR) approach. Here, feedback was limited to marking CM errors and pointing out possible resulting misconceptions. In addition, an expert map was discussed at the end of each training session, so participants had the opportunity to compare it to their own and ask questions.

(2) Participants of the T− groups did not receive any CM training but rather a control training including popular non-CM learning strategies [103,104] (see Table 1). However, the training sessions' procedure followed a similar pattern to that of the T+ groups: the instructor started with an advance organizer and introduced the learning strategy to be used that day, including metacognitive prompts. Afterwards, participants used the respective strategy to work on the learning material, whose content was identical to that of the T+ groups and related to the topic of intelligence. At

the end of the training session, an expert solution was discussed, too, allowing for comparison of own learning results and asking questions.

Table 1. Group activities during the training phase.

Training Content	Week 1	Week 2	Week 3
Scaffolding in T+ group	Skeleton map	Given set of concepts	Free practice
Learning strategy in T− group	Small group discussion	Writing a summary	Carousel workshops

T+ = CM training group (*n* = 48); T− = control training group (*n* = 45).

At the beginning of the learning phase in week 4, all participants of both training conditions received a short introduction to the CM strategy as the T− groups had not yet received any CM training. For the participants of the T+ groups, this was certainly a repetition, but it seemed important to us to refresh their knowledge at the beginning of the learning phase in order to counteract a possible diminishing effect of the training. After this short introduction, all participants filled out a questionnaire on their CM-related self-efficacy expectation. We expected the T+ groups to rate this higher than the T− groups. In order to be able to check objectively whether their assessment was accurate, they received a pre-built but error-including concept map afterwards, in which as many errors as possible should be detected and corrected within a four-minute time limit. In the following study unit, all participants read a text on the structure and functions of eukaryotic cells. Afterwards, those participants of the T+ and T− groups who were assigned to the elaboration setting were allowed to keep the text while constructing a concept map on the topic of cell biology (groups T+E and T−E), whereas those who were assigned to the retrieval setting had to construct the map without the text available (groups T+R and T−R; see Figure 1). After completing CM, all participants gave a judgment of learning on how much about the topic of cell biology they would remember one week later (see Section 2.4.2).

In the test phase (weeks 5 and 6), on the one hand, the participants' CM skills were assessed again in the context of a further construction of a concept map on the topic of cell biology, and on the other hand, we evaluated their overall learning success by application of declarative, structural, and conceptual knowledge tests (see Section 2.4.3).

2.3. Concept Map Scoring

Taking into account the expectation that participants who had the learning material available during CM (elaboration setting) would potentially specify more propositions than those who worked memory-based (retrieval setting), we used two different approaches to evaluate the quality of the concept maps constructed: an absolute (aQCM) and a balanced quality of concept map (bQCM) index. This approach has proven to be useful in our previous study [44].

(1) In order to determine the aQCM index, we followed the approach of McClure et al. [105], assigning each specified concept map proposition a value of 0 to 3 points: 0 points if there actually did not exist a relation between the concepts; 1 point if there was a relation between the concepts, but the arrow label was meaningless; 2 points if the arrow label was meaningful, but the arrow pointed in the wrong direction; 3 points if the whole proposition was correct and meaningful.

(2) In contrast, the bQCM index considers the statistically higher probability of making mistakes when specifying more propositions by putting the aQCM index into relation to the number of the participants' overall propositions specified.

Additionally, we analyzed the type of the participants' CM mistakes by classifying them as methodical (e.g., missing arrowhead) or content-related (e.g., animals—have→cell walls). On the one hand, this categorization allows for quantifying individual methodical CM skills by focusing on the concept map error ratio (number of methodical mistakes

divided by the number of overall propositions specified). On the other hand, it can be used to assess individual understanding of the learning content by focusing on the content-related error ratio (number of content-related mistakes divided by the number of overall propositions specified).

Finally, we analyzed the type of the participants' CM propositions by classifying them as recall- (R), organization- (O), or elaboration-suggesting (E). If both the relation and the associated concepts were covered completely by the learning material, this suggests an R-proposition; relations that were not explicitly named in the learning material but were constructed between concepts covered by it represent O-propositions; relations between two concepts, which included at least one that was not mentioned in the learning material, represent E-propositions, since prior knowledge needed to be integrated into the concept map (see Section 3.5).

2.4. Further Measures and Operationalizations

In the following, we describe the operationalization of the other dependent variables and some variables we controlled for baseline differences in chronological order of the consecutive study phases. Figure 2 provides an overview of all assessed variables and their relations to the study's hypotheses.

Hypothesis 1	**Hypothesis 2**	**Hypothesis 3**
If the learning material is available during CM, elaboration processes should increase.	Extensive training in CM is helpful for its successful use as a learning tool by inducing a certain familiarity especially with methodical aspects.	A combination of both factors, an extensive CM training and the availability of the learning material during CM, will improve overall learning success most.
• Types of propositions (recall-, organization-, or elaboration-suggesting) • Accuracy of propositions (aQCM and bQCM indices) • Number of content-related mistakes when creating own concept maps (content-related error ratio) • Knowledge: Declarative (reproduction of learning), structural (SJT), and conceptual (transfer of learning)	• CM-related self-efficacy expectation • Number of detected errors and accuracy of error correction when working on a pre-built but error-including concept map (error task) • Number of methodical mistakes when creating own concept maps (concept map error ratio)	• Types of propositions (recall-, organization-, or elaboration-suggesting) • Accuracy of propositions (aQCM and bQCM indices) • Knowledge: Declarative (reproduction of learning), structural (SJT), and conceptual (transfer of learning) • Judgment of learning (JOL)

Figure 2. Overview of assessed variables (bullet points) and their relations to the study's hypotheses.

2.4.1. Measures of the Training Phase (Week 1)

At the beginning of the training phase, we first assessed socio-demographic data and other relevant variables we controlled for baseline differences: (1) familiarity with CM, (2) prior knowledge of biology, and (3) reading skills (see Section 3.1).

Familiarity with CM was assessed via a questionnaire adapted from McClure et al. [105]. Using seven items, the participants rated their previous experience with CM on a five-point Likert scale (1 = *never*/*very rarely* to 5 = *very often*/*always*). The internal consistency of this scale was α = 0.88.

The participants' prior knowledge of biology was assessed in two different ways in order to obtain both a valid indicator for general knowledge of biology and a specific one regarding the topic of cell biology. For this purpose, we asked about the extent of their biology education during the past two years of schooling (*no biology* vs. *basic biology* vs. *advanced biology*) and used an 18-item knowledge test on the topic of cell biology. This test comprised (partly adapted) established items used in previous studies [24,44,106–108], and consisted of single-choice and grouping tasks. For the present sample, this test showed an internal consistency of α = 0.77 (after three items were excluded due to a lack of item-total correlation).

Since the participants had to work on the (demanding) learning texts in a given time frame during the learning units, it seemed important to us to ensure that there were no systematic variations regarding reading speed and comprehension. Therefore, we assessed

both variables using a validated instrument (Lesegeschwindigkeits- und Verständnistest [LGVT] 6–12) [109]. The task was to fulfill blank spaces by marking in each case the one meaningful out of three answer options while reading the text as fast as possible. The LGVT's internal consistency was $\alpha = 0.51$ for the present sample, indicating a lack of homogeneity of the construct [110]. Consequently, we needed to exclude the scale from further analyses.

2.4.2. Measures of the Learning Phase (Week 4)

After all participants had received the short introduction to the CM strategy at the beginning of the learning phase, they filled out a questionnaire on their CM-related self-efficacy expectation (see Section 3.2), as we expected differences between the T+ and T− groups. The scale ($\alpha = 0.83$) consisted of six items (e.g., "I feel competent in choosing the important concepts for my concept map", or "I could explain concept mapping only with words to a friend so that he could construct one on his own"), each rated on a five-point Likert scale (1 = *I do not agree at all* to 5 = *I strongly agree*). Afterwards, the participants received a pre-built but error-including concept map, in which as many errors as possible should be detected and corrected within a four-minute time limit (see Section 3.2). The pre-built concept map consisted of 18 propositions and a total of 10 common errors (e.g., missing arrowhead). Two measures were considered to obtain an objective measure of CM competence (besides the CM-related self-efficacy ratings): the number of errors detected and the accuracy of error correction.

The concept maps constructed by the participants themselves on the topic of cell biology were evaluated following the procedure described in Section 2.3 (see also Section 3.5). Interrater reliability of bQCM indexing was proven by randomly selecting one-third of the participants' concept maps for an independent re-scoring by a second rater [111]. The analysis yielded excellent interrater reliability on average ($ICC_{unjust} = 0.98$, $CI_{95\%}$ [0.93, 0.99]), indicating that participants' concept maps can be clearly judged using the coding scheme provided [112].

In addition to the CM-related measures, we assessed the judgment of learning (JOL) by asking the participants to give a metacognitive prediction regarding their learning outcome one week later (see Sections 1.2 and 3.3). As many researchers suggest that CM represents a useful metacognitive learning tool [66,113], this kind of JOL prediction has been used in several previous studies [37,38,43]. Specifically, we asked our participants to rate on a scale from 0 to 100% how much information of the learning text on cell biology they will probably remember one week later. Later on, these JOLs were compared to the corresponding objective performance measures resulting from knowledge tests of the test phase (see Section 2.4.3).

2.4.3. Measures of the Test Phase (Weeks 5 and 6)

In order to evaluate the participants' overall learning outcome, several measures were used during the two test sessions, including re-assessment of CM competence (see Section 3.5) as well as testing of declarative, structural, and conceptual knowledge (see Section 3.4). The three latter types of knowledge differ from one another: (1) declarative knowledge refers to coexisting but separate facts, (2) structural knowledge also includes relationships between such individual chunks of information, and (3) conceptual knowledge is characterized by a highly decontextualized organization of features and principles within the facts and their interrelations. Accordingly, learners with a high level of conceptual knowledge can flexibly use extensive abstract knowledge in different specific contexts [114,115].

The test phase started with the participants being given the task of constructing a concept map within 60 min by using a given set of 22 concepts taken from the text of the learning phase on the topic of cell biology. Concept map evaluation followed aQCM and bQCM indexing, as described in Section 3.3. Using the same procedure as for the corresponding measure of the learning phase, the interrater reliability was determined

for one-third of the participants' concept maps; this analysis indicated it as excellent on average ($ICC_{unjust} = 0.99$, $CI_{95\%}$ [0.90, 1.00]) [112].

After solving the CM task, participants' cell biology-related structural knowledge was assessed in a 30-minute time frame using a 55-item Similarity Judgments Test (SJT; adapted from Großschedl and Harms [116]). This SJT consisted of pairs of overall 11 cell biological concepts (e.g., "cell membrane"–"ribosomes"; "rough endoplasmic reticulum"–"proteins"), and the participants' task was to judge the semantic proximity of these pairs on a nine-point Likert scale (1 = *minimally related* to 9 = *strongly related*). In order to control for possible sequence effects, the 55 pairs were presented in a balanced manner using two test versions A and B. The participants' individual responses for each item were compared to an average rating of $n = 7$ experts. Regarding these experts' rating, the intraclass correlation coefficient indicated excellent agreement on average ($ICC_{unjust} = 0.95$, $CI_{95\%}$ [0.93, 0.97]) [112], so content validity of the procedure can be assumed. The correlation of the participants' and experts' ratings was used as indicator of structural knowledge.

The first test session (week 5) ended after a test on declarative knowledge of cell biology (60 min). This multiple-choice test comprised a total of 30 items (e.g., "Which of the following statements about the cytoplasm are correct?") and showed an internal consistency of $\alpha = 0.85$. During the last day of our study (week 6), the participants' cell biology-related conceptual knowledge was assessed. For this purpose, we have created a 15-item open answer format test ($\alpha = 0.86$) including an associated coding manual. The participants' answers were coded as either incorrect (0 points), partially correct (1 point) or completely correct (2 points). Face validity of both the test and the coding manual was checked in advance by six independent experts in the fields of biology education ($n = 3$) and test construction ($n = 3$) and found to be adequate.

3. Results

To determine the statistical correlations and differences of interest, we used parametric (Pearson correlation, *t*-test, ANOVA, and MANOVA) and non-parametric statistical analyses (Chi^2, Mann–Whitney, and Kruskal–Wallis test). The specific decision for a parametric or non-parametric method was based on distribution parameters of the analyzed variables. Accordingly, relevant empirical distributions (skewness, kurtosis, variance) were always considered within preliminary analyses in order to be able to decide whether the assumption of parametric statistical analyses were met. If there were significant deviations in one or more of these parameters, we applied nonparametric analyses accordingly.

In order to ensure the best possible clarity, the following results section is structured by content-related aspects, i.e., the analyses referring to related measures (e.g., all knowledge tests) are reported together.

3.1. Baseline Difference Testing

Since our sample was made up of self-selected participants and the number of subjects per test condition was comparatively small, we first checked for potential baseline differences between the groups on relevant variables using the data assessed at the beginning of the training phase.

Regarding the categorical variables educational level in biology and university study program, Chi^2 tests indicates an equal distribution across the four groups (see Table 2).

Regarding the metric variables age, GPA, prior knowledge in cell biology, and familiarity with CM, the results of a univariate ANOVA do not indicate any baseline differences (see Table 3).

Table 2. Chi^2 test results regarding possible baseline differences between the four (quasi-) experimental groups.

Variable	Category	Observed n in Groups				Chi^2 Test
		T+E	**T+R**	**T−E**	**T−R**	
Educational level in biology	essential	2	7	6	8	$\chi^2(6) = 7.83, p = 0.25$
	basic	17	13	9	11	
	advanced	4	5	8	4	
University study program	B. A.	12	14	8	9	$\chi^2(3) = 2.75, p = 0.43$
	B. Sc.	11	11	15	13	

T+E = CM training group (n = 23) assigned to elaboration setting (learning material available during CM); T+R = CM training group (n = 25) assigned to retrieval setting (learning material not available during CM); T−E = control training group (n = 23) assigned to elaboration setting (learning material available during CM); T−R = control training group (n = 22) assigned to retrieval setting (learning material not available during CM); an essential educational level in biology means no biology education during the past two years of schooling.

Table 3. ANOVA results regarding possible baseline differences between the four (quasi-) experimental groups.

Variable	Group								ANOVA
	T+E		**T+R**		**T−E**		**T−R**		
	M	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	
Age	22.96	7.54	21.28	4.04	22.09	5.31	21.27	4.44	$F(3, 89) = 0.50, p = 0.69$
GPA	1.95	0.67	2.06	0.74	2.06	0.67	1.80	0.53	$F(3, 89) = 0.83, p = 0.48$
Prior knowledge of cell biology	8.52	3.13	7.28	3.61	7.00	4.10	7.04	3.43	$F(3, 89) = 0.92, p = 0.44$
Familiarity with CM	1.60	0.64	1.81	0.83	1.94	0.74	2.22	0.89	$F(3, 89) = 2.49, p = 0.07$
Reading speed Reading comprehension	analyses not carried out due to the scale's lack of reliability (α = 0.51; see Section 2.4.1)								

T+E = CM training group (n = 23) assigned to elaboration setting (learning material available during CM); T+R = CM training group (n = 25) assigned to retrieval setting (learning material not available during CM); T−E = control training group (n = 23) assigned to elaboration setting (learning material available during CM); T−R = control training group (n = 22) assigned to retrieval setting (learning material not available during CM); M = mean value; SD = standard deviation; a lower GPA (=final school exam grade) indicates better performance; the maximum score for prior knowledge of cell biology was 15.

Consequently, no covariates were included additionally in subsequent statistical analyses as our results do not show any significant group differences regarding potentially confounding variables.

3.2. Concept Mapping-Related Self-Efficacy and Error Detection Task (Measures Prior to Learning Phase, Week 4)

After all participants of both training conditions received a short introduction to the CM strategy at the beginning of the learning phase in week 4, CM-related self-efficacy expectations and objective measures of CM competence by working on a pre-built but error-including concept map were assessed.

Although we had expected that the T+ groups would show higher CM-related self-efficacy expectations after the training phase than the T− groups, the corresponding Mann–Whitney test surprisingly showed no statistically significant group differences (see Table 4).

Furthermore, Mann–Whitney test could not reveal any statistically significant differences between the T+ and T− groups regarding the number of errors detected when working on a pre-built but error-including concept map. However, a comparison of the accuracy of error correction indicates a significant group difference in favor of the T+ groups, specifically with regard to the mean number of improperly corrected errors $U(n_1 = 48, n_2 = 45) = 878.0, p < 0.05, d_{Cohen} = 0.33$ (see Table 4).

Table 4. Mann–Whitney test results regarding CM-related self-efficacy and the error map task.

Variable	Group				Mann–Whitney Test
	T+		T−		
	M	*SD*	*M*	*SD*	
CM self-efficacy	5.21	1.05	5.17	1.14	*U* = 1071.0, *p* = 0.94
Error detection	7.52	2.09	7.27	1.63	*U* = 1219.5, *p* = 0.28
Proper error correction	7.48	2.12	7.27	1.63	*U* = 1209.0, *p* = 0.31
Improper error correction	0.23	0.59	0.44	0.66	*U* = 878.0, *p* < 0.05

T+ = CM training group (*n* = 48); T− = control training group (*n* = 45); *M* = mean value; *SD* = standard deviation; the maximum score for error detection and proper error correction was 10.

3.3. Metacognitive Prediction/Judgment of Learning (JOL)

An ANOVA carried out to determine differences regarding the JOL showed no significant differences between the four groups, $F(3, 89) = 0.79$, $p = 0.50$. This means, on average, all groups were equally confident about their later learning success, even though the groups nominally differed slightly in their ratings: group T−E predicted the most success ($M = 57.39$, $SD = 20.72$), followed by group T−R ($M = 55.91$, $SD = 24.23$), group T+E ($M = 53.91$, $SD = 22.10$), and group T+R ($M = 47.6$, $SD = 27.58$). This result indicates that neither an extensive CM training nor the presence or absence of the learning material during CM influence the JOLs significantly.

In addition, correlational analyses across all groups indicate statistically significant associations between these JOLs and corresponding objective performance measures resulted from the test phase's knowledge tests (see Table 5). However, the results show that the students' reference point in giving their metacognitive prediction obviously related more to declarative ($r = 0.62$, $p < 0.001$) and structural ($r = 0.52$, $p < 0.001$), but less to conceptual knowledge ($r = 0.15$, $p = 0.16$). In this context, it is noticeable that conceptual knowledge did not correlate with neither the metacognitive prediction nor with declarative or structural knowledge and thus actually seems to represent an independent domain of knowledge.

Table 5. Correlations between JOL and objective performance measures across all groups (*N* = 93).

Variable	1	2	3	4
1 Metacognitive prediction	—			
2 Declarative knowledge	0.62 ***	—		
3 Structural knowledge	0.52 ***	0.65***	—	
4 Conceptual knowledge	0.15	0.09	0.14	—

Declarative knowledge refers to coexisting but separate facts; structural knowledge also includes relationships between these facts; conceptual knowledge is characterized by a highly decontextualized organization of features and principles within the facts and their interrelations; *** = $p < 0.001$.

3.4. Declarative, Structural, and Conceptual Knowledge

In order to evaluate the students' overall learning outcome, we assessed three different types of knowledge beside CM quality in the test phase (weeks 5 and 6). To determine potential group differences between all four groups, we carried out two analyses: (1) MANOVA, taking into account declarative and structural knowledge as one latent dependent variable (linear combination), as the correlation analysis had already shown that these two types of knowledge were highly correlated in our sample, and (2) a separate ANOVA for the variable of conceptual knowledge, as it has been proven to be independent from the other types of knowledge (see Section 3.3). However, these analyses could neither reveal any significant group differences in terms of declarative and structural knowledge nor in terms of conceptual knowledge (see Table 6). On the other hand, these results are consistent with our findings above that no significant group differences could be shown regarding the metacognitive prediction of learning success in terms of knowledge acquired, even though two of the three types of knowledge were partially highly correlated across all groups.

Table 6. (M)ANOVA results regarding different types of knowledge acquired.

| Variable | Group | | | | | | | | (M)ANOVA |
| | T+E | | T+R | | T−E | | T−R | | |
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	
Declarative Knowledge	17.48	6.25	16.96	6.56	19.52	5.55	17.32	5.35	$F(6, 178) = 0.78$, $p = 0.58$
Structural Knowledge	0.40	0.18	0.41	0.19	0.41	0.20	0.39	0.21	
Conceptual Knowledge	14.61	7.26	14.20	5.63	15.43	6.36	13.27	6.34	$F(3, 89) = 0.44$, $p = 0.72$

T+E = CM training group (*n* = 23) assigned to elaboration setting (learning material available during CM); T+R = CM training group (*n* = 25) assigned to retrieval setting (learning material not available during CM); T−E = control training group (*n* = 23) assigned to elaboration setting (learning material available during CM); T−R = control training group (*n* = 22) assigned to retrieval setting (learning material not available during CM); *M* = mean value; *SD* = standard deviation; the maximum scores for declarative knowledge and conceptual knowledge were 30; the maximum score for structural knowledge was 1.0.

In addition to these analyses taking into account all four groups, we performed specific comparisons between the two settings elaboration (E) and retrieval (R), neglecting the affiliation to the T+ and T− groups, as we expected that a higher level of elaboration is associated with better learning outcomes in general, including better knowledge. However, these comparisons also did not show any significant differences in terms of significant advantages of the elaboration groups regarding the three domains of knowledge (see Table 7).

Table 7. *t*-test results regarding different types of knowledge acquired.

| Variable | Group | | | | *t*-Test |
| | E | | R | | |
	M	*SD*	*M*	*SD*	
Declarative Knowledge	18.50	5.94	17.13	5.96	$t(91) = 1.11$, $p = 0.27$
Structural Knowledge	0.40	0.19	0.40	0.20	$t(91) = 0.06$, $p = 0.95$
Conceptual Knowledge	15.02	6.76	13.77	5.93	$t(91) = 0.95$, $p = 0.34$

E = group (*n* = 46) assigned to elaboration setting (learning material available during CM); R = group (*n* = 47) assigned to retrieval setting (learning material not available during CM); *M* = mean value; *SD* = standard deviation; the maximum scores for declarative knowledge and conceptual knowledge were 30; the maximum score for structural knowledge was 1.0.

3.5. Concept Map Quality

In order to be able to evaluate the quality of the concept maps created by the participants both in the learning (week 4) and in the test phase (weeks 5 and 6), we used several indicators: (1) an absolute (aQCM) and a balanced quality of concept map (bQCM) index, (2) an analysis of the type of the participants' CM mistakes, and (3) an analysis of the type of the participants' CM propositions (see Section 2.3).

3.5.1. aQCM and bQCM Indices

Regarding the aQCM and bQCM indices, the four groups only differed in the learning phase (week 4), $F_{aQCM}(3, 88) = 4.23$, $p < 0.01$, $d_{Cohen} = 0.76$ and $\chi^2_{bQCM}(3) = 8.66$, $p < 0.05$, $d_{Cohen} = 0.52$ (see Table 8). An additional specific consideration of the learning phase's settings shows that the elaboration groups achieved significantly higher scores on average than the retrieval groups: $U_{aQCM}(n_1 = 47, n_2 = 46) = 1310.5$, $p < 0.05$, $d_{Cohen} = 0.37$ and $U_{bQCM}(n_1 = 47, n_2 = 46) = 1384.0$, $p < 0.05$, $d_{Cohen} = 0.50$. Regarding the later test phase, however, we found no significant group differences at all, indicating equal CM quality across all groups on average (see Table 8).

Table 8. ANOVA and Kruskal–Wallis test results regarding aQCM and bQCM indices.

Variable	Group								Kruskal–Wallis Test/ANOVA
	T+E		T+R		T−E		T−R		
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	
aQCM Index (LP)	106.30	37.87	78.13	35.60	124.70	59.82	113.55	50.70	$F(3, 88) = 4.23$, $p < 0.01$
bQCM Index (LP)	2.70	0.34	2.46	0.52	2.62	0.31	2.40	0.49	$\chi^2(3) = 8.66$, $p < 0.05$
aQCM Index (TP)	62.00	19.27	59.40	22.08	62.72	14.22	63.27	21.90	$F(3, 89) = 0.18$, $p = 0.91$
bQCM Index (TP)	2.40	0.44	2.50	0.38	2.37	0.40	2.32	0.38	$\chi^2(3) = 2.94$, $p = 0.40$

T+E = CM training group ($n = 23$) assigned to elaboration setting (learning material available during CM); T+R = CM training group ($n = 25$) assigned to retrieval setting (learning material not available during CM); T−E = control training group ($n = 23$) assigned to elaboration setting (learning material available during CM); T−R = control training group ($n = 22$) assigned to retrieval setting (learning material not available during CM); *M* = mean value; *SD* = standard deviation; aQCM Index = absolute quality of concept map index; bQCM = balanced quality of concept map index; LP = learning phase; TP = test phase.

3.5.2. Types of Mistakes

Additionally, we analyzed the type of the participants' CM mistakes by determining the concept map error ratio (number of methodical CM mistakes divided by the number of overall propositions specified) and the content-related error ratio (number of content-related mistakes divided by the number of overall propositions specified; see Table 9).

Table 9. Means and standard deviations regarding the type of the participants' CM mistakes.

Group		Concept Map Error Ratio (%)				Content-Related Error Ratio (%)			
		Learning Phase		Test Phase		Learning Phase		Test Phase	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Training	T+	7.0	12.0	2.0	4.0	10.0	8.0	21.0	16.0
	T−	10.0	11.0	7.0	8.0	9.0	7.0	22.0	15.0
Setting	R	10.0	14.0	4.0	6.0	12.0	9.0	21.0	15.0
	E	7.0	9.0	5.0	6.0	7.0	6.0	21.0	16.0

Means and standard deviations are reported as percentages, so an error ratio of $M = 7$ indicates that, on average, 7.0% of specified propositions were inaccurate (CM- or content-related); *M* = mean value; *SD* = standard deviation; T+ = CM training group ($n = 48$); T− = control training group ($n = 45$); R = retrieval setting (learning material not available during CM; $n = 47$); E = elaboration setting (learning material available during CM; $n = 46$).

Regarding these measures, we first performed a Mann–Whitney test to evaluate whether the CM-related mistakes differed between the participants of the two training conditions T+ and T−. An analysis of all four groups, in which the setting is also taken into account, was not advisable at this point, since it cannot be plausibly assumed that the setting influences the rate of merely methodical CM mistakes (see Sections 1.1 and 1.2). Our results show that an extensive CM training significantly reduced the CM error ratios both in the learning phase $U(n_1 = 48, n_2 = 45) = 783.5$, $p < 0.05$, $d_{\text{Cohen}} = 0.49$, and in the test phase, $U(n_1 = 48, n_2 = 45) = 656.0$, $p < 0.001$, $d_{\text{Cohen}} = 0.72$.

Afterwards, we performed a second Mann–Whitney test to evaluate whether the content-related mistakes differed between the participants of the two setting conditions elaboration (E) and retrieval (R). An analysis of all four groups, in which the training is also taken into account, was again not advisable, as this time it cannot be plausibly assumed that merely methodical CM skills influence the rate of content-related mistakes (see Section 1.3). Regarding the learning phase, our results show that the elaboration groups' content-related error ratios were significantly lower than those of the retrieval groups: $U(n_1 = 47, n_2 = 46) = 697.5$, $p < 0.01$, $d_{\text{Cohen}} = 0.64$. However, these differences go missing when looking at the test phase. Accordingly, the retrieval groups' higher content-related error ratio in the learning phase, indicating more misconceptions, does not seem to have a meaningful impact on later performance.

3.5.3. Types of Propositions

Finally, we analyzed the type of the participants' CM propositions that were specified during the learning phase in the presence or absence of the learning material. Since the total number of specified propositions differed significantly between groups, $F(3, 88) = 5.84$, $p < 0.01$, $d_{Cohen} = 0.89$, we decided to report the respective ratios of proposition categories here. A Kruskal–Wallis test taking into account all four groups showed no statistically significant differences, but non-specific trends ($p < 0.10$) regarding the E- and R-proposition ratios (see Table 10), so we decided to take a closer look at the specific differences between the settings elaboration and retrieval, neglecting the affiliation to the T+ and T− groups.

Table 10. Kruskal–Wallis test results regarding different types of CM propositions specified.

Variable	Group								Chi² Test
	T+E		T+R		T−E		T−R		
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	
R-proposition ratio (%)	93.0	10.0	87.0	16.0	88.0	10.0	84.0	14.0	$\chi^2(3) = 6.98$, $p = 0.07$
O-proposition ratio (%)	2.0	3.0	3.0	5.0	2.0	3.0	3.0	4.0	$\chi^2(3) = 1.73$, $p = 0.63$
E-proposition ratio (%)	1.0	2.0	1.0	3.0	0.4	1.0	3.0	5.0	$\chi^2(3) = 6.44$, $p = 0.09$

T+E = CM training group (*n* = 23) assigned to elaboration setting (learning material available during CM); T+R = CM training group (*n* = 25) assigned to retrieval setting (learning material not available during CM); T−E = control training group (*n* = 23) assigned to elaboration setting (learning material available during CM); T−R = control training group (*n* = 22) assigned to retrieval setting (learning material not available during CM); *M* = mean value; *SD* = standard deviation; R-proposition = recall-suggesting proposition; O-proposition = organization-suggesting proposition; E-proposition = elaboration-suggesting proposition; means and standard deviations for proposition ratios are reported as percentages, so a R-proposition ratio of *M* = 93 indicates that, on average, 93.0% of propositions in the concept map were recall-suggesting.

In respect of these potential setting effects, the respective Mann–Whitney tests indicate still no significant difference regarding R-propositions, $U(n_1 = 47, n_2 = 46) = 1259.5$, $p = 0.11$, but regarding E-propositions, $U(n_1 = 47, n_2 = 46) = 866.5$, $p < 0.05$, $d_{Cohen} = 0.35$. The latter result indicates that participants of the retrieval setting obviously specified more E-propositions ($M = 1.91$; $SD = 3.99$), although we had plausibly assumed the opposite, namely that participants who had the learning material available during CM would specify more ($M = 0.54$; $SD = 1.77$). Overall, the fact that there are significant differences between the settings, but not if the affiliation to the training groups is taken into account, can be interpreted as an indication that the presence or absence of the learning material during CM influences the quality of propositions specified considerably stronger than an extensive CM training.

4. Discussion

In our study, we tried to find out to what extent CM- and knowledge-related learning success measures differ depending on CM training intensity and the (non-) availability of learning material during the creation of concept maps in order to identify determinants for a learning-effective implementation of CM in academic contexts. Overall, our results show a significantly positive but setting-independent effect of the CM strategy training on CM-related learning outcome but not necessarily on knowledge-related learning outcome. Regarding the different setting conditions, we found both advantages and disadvantages, but in the end, the groups seem to perform equally effective, implying that CM in retrieval settings could be as effective as in elaboration settings.

4.1. Decision on Hypothesis 1

Our first hypothesis based on the assumption that non-availability of learning material during CM (R groups) could be crucial for induced cognitive processes by increasing memory-related recall [37], and thus, binding cognitive resources required for further elaboration. Accordingly, we expected an increase in elaboration processes if the learning material was available during CM (E groups), resulting in a more elaborative character

of the propositions that participants specified in their concept maps and an achievement of higher scores in subsequent knowledge tests regarding the covered learning content (see Section 1.4).

Regarding the general quality of specified propositions (aQCM and bQCM indices) and the content-related error ratios within the participants' concept maps of the learning phase, the E groups achieved higher scores on average than the R groups, but both differences levelled out until the test phase one week later (see Section 3.5.1 and Section 3.5.2). Furthermore, regarding the types of propositions, our results surprisingly show that participants of the E groups specified significantly fewer elaboration-suggesting propositions than participants of the R groups (see Section 3.5.3). Finally, regarding declarative, structural, and conceptual knowledge, our analyses did not reveal any significant advantages of the E groups (see Section 3.4). Accordingly, hypothesis 1 is not empirically supported by the data collected in our sample.

The result that, at least in the learning phase, the E groups generally specified higher quality propositions, which also offered fewer content-related mistakes, indicates that the availability of the learning material during CM obviously enables the learners to check whether their propositions specified are consistent with the text and, if necessary, to correct mistakes. Nevertheless, this process did not lead to stable learning effects in our study, as such effects should have been reflected in the test phase one week later. Accordingly, it seems likely that the availability of the learning material during a specific CM task can positively influence methodical quality aspects but does not have a beneficial effect on the consolidation of the learning content itself. This finding is in line with that of Blunt and Karpicke [37], who also consider their elaboration groups' participants to focus more on detailed representations of encoded knowledge than on improvement of cue diagnosticity, occurring more likely in retrieval practice groups. Additionally, these considerations imply a plausible explanation for the counterintuitive finding that participants of the E groups specified significantly fewer elaboration-suggesting propositions than participants of the R groups. The E groups' participants may have focused more on methodically converting the information of the text into high-quality propositions, neglecting the integration of this information into prior knowledge structures, which is crucial for the specification of propositions of a more elaborative character. Conversely, the higher proportion of elaboration processes in the R groups could obviously not support better knowledge consolidation, since the two groups achieved similar scores regarding subsequent tests of declarative, structural, and conceptual knowledge. Accordingly, this result is inconsistent with the previous findings of Blunt and Karpicke [37] as well as O'Day and Karpicke [43], stating the superiority of retrieval practice over elaboration regarding knowledge acquisition. However, the fact that the availability of the learning material does not have a significant impact on knowledge acquisition during CM is considerable as this suggests that CM in retrieval settings could be as effective as CM in elaboration settings [38,90].

4.2. Decision on Hypothesis 2

Our second hypothesis was based on the assumption that a short methodical introduction on CM principles (T− groups) is not sufficient to enable a successful use of the strategy since applying CM as an effective learning tool requires certain familiarity, especially with methodical aspects, and thus, an extensive training in CM [89,96,117,118]. Accordingly, we expected participants who took part in such a training (T+ groups) to show higher CM-related self-efficacy expectations, to achieve better results in analyzing and editing given concept maps, as well as to make fewer methodical mistakes when creating their own concept maps (see Section 1.4).

Regarding CM-related self-efficacy expectations, our results surprisingly did not show any significant group differences, i.e., participants of the T+ groups were as confident about their CM skills as those of the T− groups (see Section 3.2). Furthermore, regarding analysis and editing of given concept maps within the error task, the T+ and T− groups performed similar concerning the number of errors detected, but T+ participants corrected

these errors considerably more adequately (see Section 3.2). This result implies that the number of detected, and somehow corrected, errors is largely unaffected by the intensity of a CM training, but an extensive CM training can obviously decrease the bias toward correcting errors improperly, which in turn indicates a better understanding of methodical aspects of CM [79,88]. Finally, regarding the number of methodical mistakes made when creating own concept maps both in the learning and in the test phase, participants of the T+ groups clearly outperform those of the T− groups by showing significantly reduced the CM error ratios (see Section 3.5.2). Accordingly, the data collected in our sample partly provide empirical evidence for hypothesis 2.

The result that participants of the T+ groups were as confident about their CM skills as those of the T− groups seems surprising, but is in line with findings of our previous study [44]. The extensive training in using the demanding strategy of CM should actually have led to higher self-efficacy expectations of the T+ groups participants, especially since preliminary analyses did not reveal any significant group differences regarding experience with CM prior to our study (see Section 3.1). Accordingly, one explanation for the missing group difference could refer to a too homogeneous structure of our sample since it solely consisted of university students having many years of practical experience with learning and knowledge acquisition as well as the constant adaptation of new learning strategies. Therefore, it seems hardly surprising that all participants, regardless of their affiliation to T+ or T− groups, hold a certain degree of basic trust in their own academic abilities [119]. This consideration also implies a plausible explanation for the error task-related findings that the T+ group only deviated positively with regard to the incorrect error correction rate, whereas both groups achieved comparable results regarding error detection and proper error correction rate. If the assumption regarding the homogeneity of the sample in terms of experience in academic performance settings is correct, the aforementioned results highlight the need for a more difficult and thus better differentiating error task, including not only methodical errors but also content-related errors (see Section 2.4.2). However, the participants of the T+ group were obviously able to transfer their knowledge of methodical CM errors from the error task's more passive reception context to an active production context, which is reflected in the result that their own concept maps contained significantly fewer methodical CM mistakes both in the learning and test phase.

4.3. Decision on Hypothesis 3

Our third hypothesis consolidated hypotheses 1 and 2, as we expected that a combination of both advantageous factors, an extensive CM training and the availability of the learning material during CM (T+E), would be most effective regarding CM- and knowledge-related learning outcomes as well as their reflection in a corresponding metacognitive assessment of the participants regarding their own learning success (see Section 1.4).

However, our analyses could not reveal any significant advantages of the T+E groups compared to the other three groups (T+R, T−E, and T−R) regarding the general concept map quality (aQCM and bQCM indices; see Section 3.5.1), the types of specified propositions (E-, O-, or R-propositions; see Section 3.5.3), or the three domains of knowledge (declarative, structural, and conceptual knowledge; see Section 3.4). Consistent with this, the metacognitive predictions in terms of judgements of leaning (JOLs) between the four groups did not differ from one another (see Section 3.3). Accordingly, hypothesis 3 does not find any empirical support from the data collected in our sample.

As this hypothesis addressed a particular effectiveness of combined training and setting conditions, it is directly dependent on the two previous hypotheses. Accordingly, concerning the homogeneous performance of the four groups with regard to CM skills and knowledge measures, the same explanations apply as we have already specified in Section 4.2. In the following, we will therefore only refer to those findings that go beyond this and thus involve further implications.

The result that the aQCM and bQCM indices of the four groups differed significantly in the learning phase in favor of the T+E and T−E groups, but not any longer in the test

phase, indicates that the presence of the text could temporarily compensate a missing training on CM regarding the general quality of specified propositions, but this positive effect obviously starts fading to the same degree as the processes of forgetting occur. The assumption that significant processes of forgetting actually took place in our study is supported by the finding that the concept maps constructed by the participants of all groups in the test phase contained, on average, 2.2 times more content-related mistakes as those of the learning phase. In this context, the question arises whether a different experimental timeline (see Section 2.2) should be chosen in order to counteract such processes of forgetting. Beyond that, a potentially low level of the participants' learning motivation must be taken into account, which is particularly suggested by the results of the knowledge tests, ranking at a consistently low level of only around 40% to 60% of the achievable scores. With regard to the merely numerical scores of the knowledge test results, it seems to be slightly more advantageous for declarative knowledge acquisition if the learning material is available to the learners during CM if they have not received an extensive CM training before (T−E group). Regarding the conceptual knowledge test, it is noticeable that, on the one hand, the control training group in an elaborative setting (T−E) performed best and, on the other hand, the control training group in a retrieval setting (T−R) performed worst. The CM training groups (T+E and T+R) are in between those two, with participants in an elaboration setting performing slightly better than those in a retrieval setting. Therefore, in terms of knowledge acquisition, it might be generally better to have the learning material available during CM. If the learning material is not available, however, it seems advantageous to be familiar with CM to support knowledge acquisition during the use of the strategy as a learning tool.

Beyond that, such a familiarity with CM also seems to be beneficial in terms of metacognitive processes, since, with regard to the merely numerical scores of their judgements of learning (JOLs), the participants of the T+ groups assessed their future learning outcome slightly more accurately than those of the T− groups. In this context, it is also interesting that the participants in elaboration settings predicted a better learning outcome on average than those in retrieval settings, which confirms the results of previous studies, showing that participants in retrieval settings obviously refer to the perceived ease of memory recall of information when giving such assessments, whereas participants in elaboration settings refer to the perceived ease of information processing [37,90,91]. Nonetheless, across all four groups, the participants assessed that one week later they would only be able to remember about 50% of the information they had learned. Since the learning text on cell biology was previously assessed by experts as understandable, structured, and appropriate for the university level, it can be assumed that the JOLs of the participants were primarily quantitatively oriented towards the simple number of words. This consideration finds support if our correlational findings are also taken into account, as they suggest that the JOLs of the participants are obviously much more related to declarative and structural than to conceptual knowledge. This interesting secondary finding can most likely be attributed to the fact that our participants are simply unaccustomed to thinking in a conceptual manner, since in their everyday university life, they tend to focus on the acquisition and reproduction of declarative knowledge and sometimes structural knowledge, while conceptual knowledge, i.e., a deeper understanding of relations and dependencies between central concepts of respective fields, seems to be less required regarding completing most university study programs successfully [120], which is also reflected in popular terms such as "bulimic learning" [121].

4.4. Limitations

The previously discussed heterogeneous and sometimes unexpected findings should additionally be evaluated against the background of the study's major limitations:

(1) Our findings suggest that the structure of our sample could have been too homogeneous and thus might have undermined the occurrence of several training and setting effects. Our sample solely consisted of university students having many years of

practical experience with learning and knowledge acquisition as well as the constant adaptation of new learning strategies. In addition, it could be assumed that our student participants may not elaborate new information habitually, but are generally more familiar with rote learning, which in turn is supported by the segregated state of conceptual knowledge in our data. Therefore, we predominantly interpret the unexpectedly missing significant group differences as an effect of sampling. In this respect, it is to be expected that referring to other samples than students could reveal more distinct group differences if the participants belong to a population in which learning activities are a less central everyday topic. Furthermore, if this assumption regarding the homogeneity of our sample in terms of experience in academic performance settings is correct, our results also highlight the need for better differentiating assessments regarding few measures. In this regard, especially the error task should probably include content-related errors in addition to the methodical ones (see Section 2.4.2), as, in our previous study, we already found a similar result, indicating the task's low difficulty [44].

(2) Additionally, it should be taken into account that the final number of $N = 93$ participants was too small to reach sufficient power, ranking below 0.50 for the majority of the statistical analyses of differences performed. Therefore, a replication of our study would be desirable, taking a larger sample in order to either detect small but possibly still relevant effects or to be able to negate them without fail. Post-hoc power analyses have shown that our sample should have consisted of at least $N = 150$ participants in order to be able to detect small to medium-sized effects at a power level of 0.80 to 0.90. Considering the intensive and time-consuming care of the participants over a period of six weeks, it was not possible for us to carry out our study on a larger sample due to limited resources.

(3) Finally, the LGVT's [109] reliability was completely insufficient (<0.60) for inexplicable reasons, so we had to exclude these scales assessing reading speed and comprehension from further statistical analyses to avoid impairment of our conclusions' validity. Of course, such an exclusion is always associated with a loss of information, as we were not able to determine any baseline differences regarding reading speed and comprehension. Nevertheless, we think that this did not affect internal validity excessively, since it can be assumed that students in general offer comparable reading skills at a high level in order to be able to successfully complete their studies at all.

4.5. Prospects for Future Research

Even if our results turned out to be more heterogeneous than expected, they allow for tentative conclusions and imply suggestions for the design of prospective studies in this area that go beyond solely removing the previously discussed limitations of our study.

Regarding the setting conditions, the two groups, elaboration vs. retrieval, have both advantages and disadvantages, and in the end, they seem to perform equally effectively. This finding implies the need for taking a closer look at considerably more differentiated constellations of experimental conditions in order to uncover possible differential effects of individual parameters in the context of learning. For example, it is conceivable that corresponding effects only come into operation after a specific number of repetitions of learning phases and/or another time delay between learning and test phase, which would require respective variations and repeated short-interval follow-up measurements.

Regarding the training conditions, we were able to show that an extensive CM training has a positive effect on methodical CM skills, in particular, but not necessarily on knowledge-related learning success. This result is partly inconsistent with the results of our previous study, in which we found both significantly better CM skills and knowledge-related advantages of participants who received an extensive training [44]. This missing effect of an extensive training on knowledge acquisition in our present study allows for drawing two possible conclusions:

(1) Possibly, the 30-minute short introduction of the T− groups regarding to the most important CM principles was just as effective as the extensive CM training of the T+ group. Such an effect could be attributed in two different ways. Either it again must be interpreted against the background of our sample of experienced and generally successful learners (which seems unlikely, since the overall performance of the groups ranged on a comparatively low level), or it is due to a generally low level of learning motivation, since incentives (e.g., study credits) for participation in our study were not offered to the students (see Section 2.1). If the latter should be true, it would be necessary to consider the participants' learning motivation as a covariate in future studies.

(2) Additionally, it is once more possible that resounding effects of an extensive CM training could be more successfully activated by a different experimental timeline than the one we chose (see Section 2.2). For example, it is conceivable that the training's success differs depending on the overall number of repetitions and/or the duration of individual training sessions and/or the time interval between these sessions, which could be clarified in future studies by systematic variation.

Finally, it would be interesting to investigate the potential effect of implementing an effective training with alternating phases of text availability vs. non-availability on overall learning success. In the studies of Karpicke and Blunt [38], Blunt and Karpicke [37], and O'Day and Karpicke [43], such an alternation was implemented, but the participants received only a short introduction to CM (~T− condition in our study) and furthermore, these studies' textual learning material consisted of approximately only 300 words, so ecological validity in terms of appropriateness regarding university level can be questioned. Future studies could, therefore, focus on experimental conditions that are closer to university learning and performance contexts, as we did in our study.

The design and implementation of such studies would help identify the conditions for an efficient use of CM and thus provide learners with an effective strategy that can support a deeper understanding of a field.

Institutional Review Board Statement: Our study was approved by the German Research Foundation (DFG; project number: GR 4763/2-1). Before participation, all subjects received a detailed written subject information in accordance with the current ethical guidelines laid down by the University of Cologne (Germany) and the German Psychological Society [122], including the following information: aims and course of the investigation, absolute voluntariness of participation, possibility of dropping out of participation at any time, guaranteed protection of data privacy (collection of only anonymized data), possibility of requesting data cancelation at any time, no-risk character of study participation, and contact information in case of any questions or problems. Written informed consent was obtained from all participants prior to the study, according to the current version of the Declaration of Helsinki [123,124]. Anonymity was ensured by using individual codes that the students created themselves. Data storage meets current European data protection regulations [125].

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Conradty, C.; Sotiriou, S.A.; Bogner, F.X. How Creativity in STEAM Modules Intervenes with Self-Efficacy and Motivation. *Educ. Sci.* **2020**, *10*, 70. [CrossRef]
2. Mulnix, A.B.; Vandegrift, E.V.H. A Tipping Point in STEM Education Reform. *J. Coll. Sci. Teach.* **2014**, *43*, 14–16. [CrossRef]
3. Tripto, J.; Assaraf, O.B.Z.; Snapir, Z.; Amit, M. How is the body's systemic nature manifested amongst high school biology students? *Instr. Sci.* **2017**, *45*, 73–98. [CrossRef]
4. Udovic, D.; Morris, D.; Dickman, A.; Postlethwait, J.; Wetherwax, P. Workshop Biology: Demonstrating the Effectiveness of Active Learning in an Introductory Biology Course. *BioScience* **2002**, *52*, 272–281. [CrossRef]
5. Arwood, L. Teaching Cell Biology to Nonscience Majors Through Forensics, Or How to Design a Killer Course. *Cell Biol. Educ.* **2004**, *3*, 131–138. [CrossRef]
6. Jesionkowska, J.; Wild, F.; Deval, Y. Active Learning Augmented Reality for STEAM Education—A Case Study. *Educ. Sci.* **2020**, *10*, 198. [CrossRef]
7. Labov, J.B.; Reid, A.H.; Yamamoto, K.R. Integrated Biology and Undergraduate Science Education: A New Biology Education for the Twenty-First Century? *CBE Life Sci. Educ.* **2010**, *9*, 10–16. [CrossRef]
8. Lathwesen, C.; Belova, N. Escape Rooms in STEM Teaching and Learning—Prospective Field or Declining Trend? A Literature Review. *Educ. Sci.* **2021**, *11*, 308. [CrossRef]
9. Silveira, L.F.; Xavier, C.S.; Perez, M.A.; Bozza, D.A.; Furtado-Alle, L.; Hass, I.; Tureck, L.V. Unravelling the Central Dogma of Biology in an active way: A case study. *J. Biol. Educ.* **2021**, *55*, 1–14. [CrossRef]
10. Swaid, S.I. Bringing Computational Thinking to STEM Education. *Procedia Manuf.* **2015**, *3*, 3657–3662. [CrossRef]
11. Young, P.W. Corrigendum: Student-Produced Video of Role-Plays on Topics in Cell Biology and Biochemistry: A Novel Undergraduate Group Work Exercise. *Front. Educ.* **2020**, *5*, 115. [CrossRef]
12. McGreevy, K.M.; Church, F.C. Active Learning: Subtypes, Intra-Exam Comparison, and Student Survey in an Undergraduate Biology Course. *Educ. Sci.* **2020**, *10*, 185. [CrossRef]
13. Aguilera, D.; Ortiz-Revilla, J. STEM vs. STEAM Education and Student Creativity: A Systematic Literature Review. *Educ. Sci.* **2021**, *11*, 331. [CrossRef]
14. Preszler, R. Cooperative Concept Mapping: Improving Performance in Undergraduate Biology. *J. Coll. Sci. Teach.* **2004**, *33*, 30–35.
15. Tullis, J.G.; Goldstone, R.L. Instruction in Computer Modeling Can Support Broad Application of Complex Systems Knowledge. *Front. Educ.* **2017**, *2*, 4. [CrossRef]
16. Verhoeff, R.P.; Knippels, M.-C.P.J.; Gilissen, M.G.R.; Boersma, K.T. The Theoretical Nature of Systems Thinking. Perspectives on Systems Thinking in Biology Education. *Front. Educ.* **2018**, *3*, 40. [CrossRef]
17. Campbell, N.A.; Urry, L.A.; Cain, M.L.; Wasserman, S.A.; Minorsky, P.V.; Orr, R.B. *Biology: A Global Approach*, 12th ed.; Pearson Education Limited: Harlow, UK, 2021; ISBN 978-1292341637.
18. Barak, J.; Sheva, B.; Gorodetsky, M.; Gurion, B. As 'process' as it can get: Students' understanding of biological processes. *Int. J. Sci. Educ.* **1999**, *21*, 1281–1292. [CrossRef]
19. Brandstädter, K.; Harms, U.; Großschedl, J. Assessing System Thinking through Different Concept-Mapping Practices. *Int. J. Sci. Educ.* **2012**, *34*, 2147–2170. [CrossRef]
20. Garvin-Doxas, K.; Klymkowsky, M.; Elrod, S. Building, Using, And Maximizing the Impact of Concept Inventories in the Biological Sciences: Report on a National Science Foundation-sponsored Conference on the Construction of Concept Inventories in the Biological Sciences. *CBE Life Sci. Educ.* **2007**, *6*, 277–282. [CrossRef]
21. Grotzer, T.A.; Solis, S.L.; Tutwiler, M.S.; Cuzzolino, M.P. A study of students' reasoning about probabilistic causality: Implications for understanding complex systems and for instructional design. *Instr. Sci.* **2017**, *45*, 25–52. [CrossRef]
22. Haslam, F.; Treagust, D.F. Diagnosing secondary students' misconceptions of photosynthesis and respiration in plants using a two-tier multiple choice instrument. *J. Biol. Educ.* **1987**, *21*, 203–211. [CrossRef]
23. Hmelo-Silver, C.E.; Azevedo, R. Understanding Complex Systems: Some Core Challenges. *J. Learn. Sci.* **2006**, *15*, 53–61. [CrossRef]
24. Shi, J.; Wood, W.B.; Martin, J.M.; Guild, N.A.; Vicens, Q.; Knight, J.K. A Diagnostic Assessment for Introductory Molecular and Cell Biology. *CBE Life Sci. Educ.* **2010**, *9*, 453–461. [CrossRef]
25. Mandl, H.; Friedrich, H.F. *Handbuch Lernstrategien [Handbook of Learning Strategies]*; Hogrefe: Göttingen, Germany, 2006; ISBN 978-3801718138.
26. Anderson, J.R. *The Architecture of Cognition*; Lawrence Erlbaum Associates, Inc.: Hillsdale, NJ, USA, 1983; ISBN 978-0674044258.
27. Weinstein, C.E.; Mayer, R.E. The teaching of learning strategies. In *Handbook of Research on Teaching*, 3rd ed.; Wittrock, M.C., Ed.; Macmillan: New York, NY, USA, 1986; pp. 315–327.
28. Kalyuga, S. Knowledge elaboration: A cognitive load perspective. *Learn. Instr.* **2009**, *19*, 402–410. [CrossRef]
29. Wild, K.-P.; Schiefele, U. Lernstrategien im Studium: Ergebnisse zur Faktorenstruktur und Reliabilität eines neuen Fragebogens [Learning strategies in academic studies: Results on factorial structure and reliability of a new questionnaire]. *J. Individ. Differ.* **1994**, *15*, 185–200.

30. Wadouh, J. Vernetzung und Kumulatives Lernen im Biologieunterricht der Gymnasialklasse 9 [Cross-Linked Thinking and Cumulative Learning in Grade 9 Secondary School Biology Classes]. Ph.D. Thesis, University Duisburg-Essen, Duisburg, Germany, 2007.

31. Carpenter, S.K.; Pashler, H.; Cepeda, N.J. Using tests to enhance 8th grade students' retention of U.S. history facts. *Appl. Cogn. Psychol.* **2009**, *23*, 760–771. [CrossRef]

32. Karpicke, J.D. Retrieval-based learning: A decade of progress. In *Cognitive Psychology of Memory*, 2nd ed.; Wixted, J.T., Ed.; Academic Press: Oxford, UK, 2017; pp. 487–514.

33. Karpicke, J.D.; Roediger, H.L. The Critical Importance of Retrieval for Learning. *Science* **2008**, *319*, 966–968. [CrossRef] [PubMed]

34. Pyc, M.A.; Rawson, K.A. Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *J. Mem. Lang.* **2009**, *60*, 437–447. [CrossRef]

35. Roediger, H.L.; Karpicke, J.D. The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspect. Psychol. Sci.* **2006**, *1*, 181–210. [CrossRef] [PubMed]

36. Großschedl, J.; Tröbst, S. Biologie lernen durch Concept Mapping: Bedeutung eines Lernstrategietrainings für kognitive Belastung, kognitive Prozesse und Lernleistung—Kurzdarstellung des DFG–Projekts [Learning biology using concept mapping: Importance of a learning strategy training for cognitive load, cognitive processes and learning performance—Brief description of the DFG project]. *Zeitschrift für Didaktik der Biologie (ZDB)-Biologie Lehren und Lernen* **2018**, *22*, 20–30. [CrossRef]

37. Blunt, J.R.; Karpicke, J.D. Learning with retrieval-based concept mapping. *J. Educ. Psychol.* **2014**, *106*, 849–858. [CrossRef]

38. Karpicke, J.D.; Blunt, J.R. Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping. *Science* **2011**, *331*, 772–775. [CrossRef]

39. Karpicke, J.D.; Smith, M.A. Separate mnemonic effects of retrieval practice and elaborative encoding. *J. Mem. Lang.* **2012**, *67*, 17–29. [CrossRef]

40. Brame, C.J.; Biel, R. Test-Enhanced Learning: The Potential for Testing to Promote Greater Learning in Undergraduate Science Courses. *CBE Life Sci. Educ.* **2015**, *14*, es4. [CrossRef] [PubMed]

41. Karpicke, J.D. Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *J. Exp. Psychol. Gen.* **2009**, *138*, 469–486. [CrossRef]

42. Moreira, B.F.T.; Pinto, T.S.S.; Starling, D.S.V.; Jaeger, A. Retrieval Practice in Classroom Settings: A Review of Applied Research. *Front. Educ.* **2019**, *4*, 5. [CrossRef]

43. O'Day, G.M.; Karpicke, J.D. Comparing and combining retrieval practice and concept mapping. *J. Educ. Psychol.* **2021**, *113*, 986–997. [CrossRef]

44. Becker, L.B.; Welter, V.D.E.; Aschermann, E.; Großschedl, J. Comprehension-Oriented Learning of Cell Biology: Do Different Training Conditions Affect Students' Learning Success Differentially? *Educ. Sci.* **2021**, *11*, 438. [CrossRef]

45. Morse, D.; Jutras, F. Implementing Concept-based Learning in a Large Undergraduate Classroom. *CBE Life Sci. Educ.* **2008**, *7*, 243–253. [CrossRef] [PubMed]

46. Okebukola, P.A. Attaining meaningful learning of concepts in genetics and ecology: An examination of the potency of the concept-mapping technique. *J. Res. Sci. Teach.* **1990**, *27*, 493–504. [CrossRef]

47. Okebukola, P.A.; Jegede, O.J. Cognitive preference and learning mode as determinants of meaningful learning through concept mapping. *Sci. Educ.* **1988**, *72*, 489–500. [CrossRef]

48. Pruett, J.L.; Weigel, E.G. Concept Map Assessment Reveals Short-Term Community-Engaged Fieldwork Enhances Sustainability Knowledge. *CBE Life Sci. Educ.* **2020**, *19*, ar38. [CrossRef] [PubMed]

49. Hegarty-Hazel, E.; Prosser, M. Relationship between Students' Conceptual Knowledge and Study Strategies—Part 2: Student Learning in Biology. *Int. J. Sci. Educ.* **1991**, *13*, 303–312. [CrossRef]

50. Martínez, G.; Pérez, Á.L.; Suero, M.I.; Pardo, P.J. The Effectiveness of Concept Maps in Teaching Physics Concepts Applied to Engineering Education: Experimental Comparison of the Amount of Learning Achieved with and without Concept Maps. *J. Sci. Educ. Technol.* **2013**, *22*, 204–214. [CrossRef]

51. Gahr, A.A. Cooperative Chemistry: Concept Mapping in the Organic Chemistry Lab. *J. Coll. Sci. Teach.* **2003**, *32*, 311.

52. Jack, G.U. Concept Mapping and Guided Inquiry as Effective Techniques for Teaching Difficult Concepts in Chemistry: Effect on Students' Academic Achievement. *J. Educ. Pract.* **2013**, *4*, 9–15.

53. Cañas, A.J.; Novak, J.D. Concept Mapping Using CmapTools to Enhance Meaningful Learning. In *Knowledge Cartography: Software Tools and Mapping Techniques*, 2nd ed.; Okada, A., Buckingham Shum, S.J., Sherborne, T., Eds.; Springer: London, UK, 2014; pp. 23–45.

54. Cronin, P.J.; Dekhers, J.; Dunn, J.G. A procedure for using and evaluating concept maps. *Res. Sci. Educ.* **1982**, *12*, 17–24. [CrossRef]

55. Derbentseva, N.; Safayeni, F.; Cañas, A.J. Concept maps: Experiments on dynamic thinking. *J. Res. Sci. Teach.* **2007**, *44*, 448–465. [CrossRef]

56. Novak, J.D. Concept mapping: A useful tool for science education. *J. Res. Sci. Teach.* **1990**, *27*, 937–949. [CrossRef]

57. Novak, J.D. *Learning, Creating, and Using Knowledge: Concept Maps as Facilitative Tools in Schools and Corporations*, 2nd ed.; Routledge: New York, NY, USA, 2010; ISBN 978-0415991858.

58. Novak, J.D.; Cañas, A. *The Theory Underlying Concept Maps and How to Construct Them*; Florida Institute for Human and Machine Cognition: Pensacola, FL, USA, 2008; Available online: http://cmap.ihmc.us/docs/pdf/TheoryUnderlyingConceptMaps.pdf (accessed on 21 August 2021).

59. Quillian, M.R. Semantic networks. In *Semantic Information Processing*; Minsky, M.L., Ed.; MIT Press: Cambridge, MA, USA, 1968; pp. 227–270.

60. Collins, A.M.; Quillian, M.R. Retrieval time from semantic memory. *J. Verbal Learn. Verbal Behav.* **1969**, *8*, 240–247. [CrossRef]

61. Buchner, A.; Brandt, M. Gedächtniskonzeptionen und Wissensrepräsentationen [Conceptions of memory and representations of knowledge]. In *Allgemeine Psychologie [General Psychology]*; Müsseler, J., Rieger, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2017; pp. 401–434.

62. Renkl, A.; Nückles, M. Lernstrategien der externen Visualisierung [Learning strategies of external visualization]. In *Handbuch Lernstrategien [Handbook of Learning Strategies]*; Mandl, H., Friedrich, H.F., Eds.; Hogrefe: Göttingen, Germany, 2006; pp. 135–147.

63. den Elzen-Rump, V.; Leutner, D. Naturwissenschaftliche Sachtexte verstehen—Ein computerbasiertes Trainingsprogramm für Schüler der 10. Jahrgangsstufe zum selbstregulierten Lernen mit einer Mapping-Strategie [Understanding scientific texts—A computer-based training program for 10th grade students for self-regulated learning with a mapping strategy]. In *Selbstregulation Erfolgreich Fördern [Promoting Self-Regulation Successfully]*; Landmann, M., Schmitz, B., Eds.; Kohlhammer: Stuttgart, Germany, 2007; pp. 251–268.

64. Jüngst, K.L. Studien zur didaktischen Nutzung von Concept Maps [Studies on the didactic use of concept maps]. *Unterrichtswissenschaft* **1995**, *3*, 229–250.

65. Jüngst, K.L.; Strittmatter, P. Wissensstrukturdarstellung: Theoretische Ansätze und praktische Relevanz [Representation of knowledge structures: Theoretical approaches and practical relevance]. *Unterrichtswissenschaft* **1995**, *23*, 194–207.

66. Cañas, A.; Coffey, J.; Carnot, M.J.; Feltovich, P.; Hoffman, R.R.; Feltovich, J.; Novak, J.D. A Summary of Literature Pertaining to the Use of Concept Mapping Techniques and Technologies for Education and Performance Support. Available online: https://www.ihmc.us/users/acanas/Publications/ConceptMapLitReview/IHMC%20Literature%20Review%20on%20Concept%20Mapping.pdf (accessed on 21 August 2021).

67. Cadorin, L.; Bagnasco, A.; Rocco, G.; Sasso, L. An integrative review of the characteristics of meaningful learning in healthcare professionals to enlighten educational practices in health care. *Nurs. Open* **2014**, *1*, 3–14. [CrossRef] [PubMed]

68. Hattie, J. *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*, 1st ed.; Routledge: New York, NY, USA, 2009; ISBN 978-0415476171.

69. Haugwitz, M. Kontextorientiertes Lernen und Concept Mapping im Fach Biologie [Context-Oriented Learning and Concept Mapping in Biology]. Ph.D. Thesis, University Duisburg-Essen, Duisburg, Germany, 2009.

70. Hilbert, T.S.; Renkl, A. Concept mapping as a follow-up strategy to learning from texts: What characterizes good and poor mappers? *Instr. Sci.* **2008**, *36*, 53–73. [CrossRef]

71. Nesbit, J.C.; Adesope, O.O. Learning with Concept and Knowledge Maps: A Meta-Analysis. *Rev. Educ. Res.* **2006**, *76*, 413–448. [CrossRef]

72. Mayer, R.E. Rote Versus Meaningful Learning. *Theory Pract.* **2002**, *41*, 226–232. [CrossRef]

73. Nückles, M.; Hübner, S.; Dümer, S.; Renkl, A. Expertise reversal effects in writing-to-learn. *Instr. Sci.* **2010**, *38*, 237–258. [CrossRef]

74. Bramwell-Lalor, S.; Rainford, M. The Effects of Using Concept Mapping for Improving Advanced Level Biology Students' Lower- and Higher-Order Cognitive Skills. *Int. J. Sci. Educ.* **2014**, *36*, 839–864. [CrossRef]

75. Kinchin, I.M. Visualising knowledge structures in biology: Discipline, curriculum and student understanding. *J. Biol. Educ.* **2011**, *45*, 183–189. [CrossRef]

76. Mintzes, J.J.; Wandersee, J.H.; Novak, J.D. Meaningful learning in science: The human constructivist perspective. In *Handbook of Academic Learning: Construction of Knowledge*; Phye, G.D., Ed.; Academic Press: San Diego, CA, USA, 1997; pp. 405–447.

77. Ritchhart, R.; Turner, T.; Hadar, L. Uncovering students' thinking about thinking using concept maps. *Metacogn. Learn.* **2009**, *4*, 145–159. [CrossRef]

78. Salmon, D.; Kelly, M. *Using Concept Mapping to Foster Adaptive Expertise: Enhancing Teacher Metacognitive Learning to Improve Student Academic Performance*; Peter Lang: New York, NY, USA, 2014; ISBN 978-1433122705.

79. Kapuza, A. How Concept Maps with and without a List of Concepts Differ: The Case of Statistics. *Educ. Sci.* **2020**, *10*, 91. [CrossRef]

80. Ritter, R.; Wehner, A.; Lohaus, G.; Krämer, P. Pre-service Teachers' Beliefs about Inclusive Education before and after Multi-Compared to Mono-professional Co-teaching: An Exploratory Study. *Front. Educ.* **2019**, *4*, 101. [CrossRef]

81. Marzetta, K.; Mason, H.; Wee, B. 'Sometimes They Are Fun and Sometimes They Are Not': Concept Mapping with English Language Acquisition (ELA) and Gifted/Talented (GT) Elementary Students Learning Science and Sustainability. *Educ. Sci.* **2018**, *8*, 13. [CrossRef]

82. Andrews, K.E.; Tressler, K.D.; Mintzes, J.J. Assessing environmental understanding: An application of the concept mapping strategy. *Environ. Educ. Res.* **2008**, *14*, 519–536. [CrossRef]

83. Martin, B.L.; Mintzes, J.J.; Clavijo, I.E. Restructuring knowledge in Biology: Cognitive processes and metacognitive reflections. *Int. J. Sci. Educ.* **2000**, *22*, 303–323. [CrossRef]

84. Mintzes, J.J.; Wandersee, J.H.; Novak, J.D. Assessing understanding in biology. *J. Biol. Educ.* **2001**, *35*, 118–124. [CrossRef]

85. Pearsall, N.R.; Skipper, J.E.J.; Mintzes, J.J. Knowledge restructuring in the life sciences: A longitudinal study of conceptual change in biology. *Sci. Educ.* **1997**, *81*, 193–215. [CrossRef]

86. Quinn, H.J.; Mintzes, J.J.; Laws, R.A. Successive Concept Mapping: Assessing Understanding in College Science Classes. *J. Coll. Sci. Teach.* **2003**, *33*, 12–16.

87. Hilbert, T.; Nückles, M.; Matzel, S. Concept mapping for learning from text: Evidence for a worked-out-map-effect. In Proceedings of the 8th International Conference for the Learning Sciences, Utrecht, The Netherlands, 23–28 June 2008; International Society of the Learning Sciences: Utrecht, The Netherlands, 2008; Volume 1, pp. 358–365.

88. Sumfleth, E.; Neuroth, J.; Leutner, D. Concept Mapping—eine Lernstrategie muss man lernen [Concept mapping—A learning strategy has to be learned]. *CHEMKON* **2010**, *17*, 66–70. [CrossRef]

89. Mintzes, J.J.; Canas, A.; Coffey, J.; Gorman, J.; Gurley, L.; Hoffman, R.; McGuire, S.Y.; Miller, N.; Moon, B.; Trifone, J.; et al. Comment on "Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping". *Science* **2011**, *334*, 453. [CrossRef]

90. Karpicke, J.D. Retrieval-Based Learning:Active Retrieval Promotes Meaningful Learning. *Curr. Dir. Psychol. Sci.* **2012**, *21*, 157–163. [CrossRef]

91. Roediger, H.L.; Karpicke, J.D. Test-Enhanced Learning:Taking Memory Tests Improves Long-Term Retention. *Psychol. Sci.* **2006**, *17*, 249–255. [CrossRef] [PubMed]

92. Karpicke, J.D.; Grimaldi, P.J. Retrieval-Based Learning: A Perspective for Enhancing Meaningful Learning. *Educ. Psychol. Rev.* **2012**, *24*, 401–418. [CrossRef]

93. Koriat, A. Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *J. Exp. Psychol. Gen.* **1997**, *126*, 349–370. [CrossRef]

94. Koriat, A. Metacognition and consciousness. In *Cambridge Handbook of Consciousness*; Zelazo, P.D., Moscovitch, M., Thompson, E., Eds.; Cambridge University Press: New York, NY, USA, 2007; pp. 289–325.

95. Ajaja, O.P. Concept mapping as a study skill. *Int. J. Educ. Sci.* **2011**, *3*, 49–57. [CrossRef]

96. Schwendimann, B. Concept Mapping. In *Encyclopedia of Science Education*; Gunstone, R., Ed.; Springer: Dordrecht, The Netherlands, 2015; pp. 198–202.

97. Jonassen, D.H.; Beissner, K.; Yacci, M. *Structural Knowledge: Techniques for Representing, Conveying, And Acquiring Structural Knowledge*; Lawrence Erlbaum Associates, Inc.: Hillsdale, NJ, USA, 1993; ISBN 978-0805810097.

98. Chularut, P.; DeBacker, T.K. The influence of concept mapping on achievement, self-regulation, and self-efficacy in students of English as a second language. *Contemp. Educ. Psychol.* **2004**, *29*, 248–263. [CrossRef]

99. Arnaudin, M.W.; Mintzes, J.J. Students' alternative conceptions of the human circulatory system: A cross-age study. *Sci. Educ.* **1985**, *69*, 721–733. [CrossRef]

100. Hilbert, T.S.; Renkl, A. Learning how to use a computer-based concept-mapping tool: Self-explaining examples helps. *Comput. Human Behav.* **2009**, *25*, 267–274. [CrossRef]

101. Großschedl, J.; Harms, U. Effekte metakognitiver Prompts auf den Wissenserwerb beim Concept Mapping und Notizen Erstellen [Effects of metacognitive prompts on knowledge acquisition in concept mapping and note taking]. *Z. Didakt. Nat.* **2013**, *19*, 375–395.

102. Großschedl, J.; Harms, U. Metakognition—Denken aus der Vogelperspektive [Metacognition—Thinking from a bird's eye view]. In *Biologie Methodik [Methodology of Biology]*, 4th ed.; Spörhase-Eichmann, U., Ruppert, W., Eds.; Cornelsen: Berlin, Germany, 2018; pp. 48–52.

103. Ajaja, O.P. Which way do we go in biology teaching? Lecturing, concept mapping, cooperative learning or learning cycle? *Electron. J. Sci. Educ.* **2013**, *17*, 1–37. [CrossRef]

104. Reader, W.; Hammond, N. Computer-based tools to support learning from hypertext: Concept mapping tools and beyond. *Comput. Educ.* **1994**, *22*, 99–106. [CrossRef]

105. McClure, J.R.; Sonak, B.; Suen, H.K. Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. *J. Res. Sci. Teach.* **1999**, *36*, 475–492. [CrossRef]

106. Champagne Queloz, A.; Klymkowsky, M.W.; Stern, E.; Hafen, E.; Köhler, K. Diagnostic of students' misconceptions using the Biological Concepts Instrument (BCI): A method for conducting an educational needs assessment. *PLoS ONE* **2017**, *12*, e0176906. [CrossRef] [PubMed]

107. Großschedl, J.; Mahler, D.; Harms, U. Construction and Evaluation of an Instrument to Measure Content Knowledge in Biology: The CK-IBI. *Educ. Sci.* **2018**, *8*, 145. [CrossRef]

108. Kleickmann, T.; Großschedl, J.; Harms, U.; Heinze, A.; Herzog, S.; Hohenstein, F.; Köller, O.; Kröger, J.; Lindmeier, A.; Loch, C. Professionswissen von Lehramtsstudierenden der mathematisch-naturwissenschaftlichen Fächer—Testentwicklung im Rahmen des Projekts KiL [Professional knowledge of student teachers of mathematical and scientific subjects—Test development within the KiL project]. *Unterrichtswissenschaft* **2014**, *42*, 280–288.

109. Schneider, W.; Schlagmüller, M.; Ennemoser, M. *Lesegeschwindigkeits-und Verständnistest für die Klassen 6–12 [Reading Speed and Comprehension Test for Grades 6–12]*; Hogrefe: Göttingen, Germany, 2007; ISBN 978-3403066385.

110. Döring, N.; Bortz, J.; Poeschl-Guenther, S. *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften [Research Methods and Evaluation in the Social and Human Sciences]*, 5th ed.; Springer: Berlin/Heidelberg, Germany, 2016.

111. Wirtz, M.; Caspar, F. *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen [Interrater Consistency and Interrater Reliability: Methods for Determining and Improving the Reliability of Assessments Using Observational Systems and Rating Scales]*, 1st ed.; Hogrefe: Göttingen, Germany, 2002; ISBN 978-3801716462.

112. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef]

113. Townsend, C.L.; Heit, E. Judgments of learning and improvement. *Mem. Cognit.* **2011**, *39*, 204–216. [CrossRef]

114. Anderson, J.R. *Cognitive Psychology and Its Implications*, 9th ed.; Worth Publishers: New York, NY, USA, 2020; ISBN 978-1319067113.

115. Watson, M.K.; Pelkey, J.; Noyes, C.R.; Rodgers, M.O. Assessing Conceptual Knowledge Using Three Concept Map Scoring Methods. *J. Eng. Educ.* **2016**, *105*, 118–146. [CrossRef]

116. Großschedl, J.; Harms, U. Assessing conceptual knowledge using similarity judgments. *Stud. Educ. Eval.* **2013**, *39*, 71–81. [CrossRef]

117. Roessger, K.M.; Daley, B.J.; Hafez, D.A. Effects of teaching concept mapping using practice, feedback, and relational framing. *Learn. Instr.* **2018**, *54*, 11–21. [CrossRef]

118. Schroeder, N.L.; Nesbit, J.; Anguiano, C.J.; Adesope, O.O. Studying and Constructing Concept Maps: A Meta-Analysis. *Educ. Psychol. Rev.* **2018**, *30*, 431–455. [CrossRef]

119. Sears, D.O. College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *J. Pers. Soc. Psychol.* **1986**, *51*, 515–530. [CrossRef]

120. Erpenbeck, J.; Sauter, W. *Stoppt die Kompetenzkatastrophe! Wege in eine neue Bildungswelt [Stop the Competence Disaster! Paths to A New World of Education]*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2019; ISBN 978-3662596777.

121. Wild, E.; Esdar, W. Beiträge der Pädagogischen Psychologie zur Wissenschaftspolitik [Contributions of educational psychology to science policy]. In *Handbuch Wissenschaftspolitik [Handbook Science Policy]*, 2nd ed.; Simon, D., Knie, A., Hornbostel, S., Zimmermann, K., Eds.; Springer: Wiesbaden, Germany, 2016; pp. 191–205.

122. Berufsethische Richtlinien des Berufsverbandes Deutscher Psychologinnen und Psychologen e. V. und der Deutschen Gesellschaft für Psychologie [Professional Ethical Guidelines of the Professional Association of German Psychologists e. V. and the German Psychological Society]. Available online: https://www.bdp-verband.de/binaries/content/assets/beruf/ber-foederation-2016 .pdf (accessed on 21 August 2021).

123. World Medical Association. WMA's Declaration of Helsinki Serves as Guide to Physicians. *J. Am. Med. Assoc.* **1964**, *189*, 33–34. [CrossRef]

124. World Medical Association. Declaration of Helsinki. Ethical Principles for Medical Research Involving Human Subjects. *J. Am. Med. Assoc.* **2013**, *310*, 2191–2194. [CrossRef] [PubMed]

125. European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Off. J. Eur. Union* **2016**, *59*, 294.