Metacognition and Learning (2021) 16:233–253
https://doi.org/10.1007/s11409-020-09246-4

Effects of keyword tasks and biasing titles on metacognitive monitoring and recall

Marie Lippmann[1] · Robert W. Danielson[2] · Neil H. Schwartz[1] · Hermann Körndle[3] · Susanne Narciss[3]

Received: 8 July 2019 / Accepted: 28 September 2020 / Published online: 12 November 2020

Abstract
This investigation examines the effects of keyword tasks (Immediate vs. Delayed) on metacognitive monitoring, study regulation, and recall in multi-step learning tasks, which require learning information from expository texts. The titles of the expository texts were biased towards information that was either stated close to the title (Related/Close), distant from the title (Related/Distant), or unrelated to the title (Unrelated). Based on the Cue-Utilization Framework, we hypothesized that learners' metacognitive monitoring and study regulation would be informed by mnemonic cues derived from text-titles and keyword tasks. Two hundred and thirteen American undergraduate students studied six expository texts, generated keywords, provided judgments of learning, and wrote about what they recalled before and after a self-regulated restudy trial. In line with our main hypothesis, the results revealed that learners who generated keywords immediately overestimated their current state of learning to a greater extent than learners who generated keywords with a delay. Contrary to our expectations, the greater monitoring accuracy observed in the delayed keyword group did not result in more effective restudy behavior. Learners in both keyword groups were able to improve their recall performance from their first to their second set of recall tasks, but interestingly, only learners in the immediate keyword group utilized the restudy trial to close knowledge gaps between information, which was stated close to versus distant from the title.

Keywords Recall · Judgments of learning · Self-regulated learning · Metacognitive monitoring

Metacognitive monitoring and study regulation

Imagine Marie, a student who is studying for an upcoming final in her undergraduate psychology class. Wanting to do well on the final, she will likely spend hours, spread out across multiple days, studying the course materials. She may re-read some of the chapters the

✉ Marie Lippmann
mlippmann@csuchico.edu

Extended author information available on the last page of the article

professor had assigned, re-read some notes from the class, or even re-watch some of the lectures if they are available. Throughout this process, Marie will monitor her current understanding of the material and compare this state of learning to the level of knowledge that is expected on the upcoming final. This process of self-assessment is metacognitive in nature since Marie is monitoring her own learning, based on her self-assessment of her own knowledge (Azevedo 2009; Azevedo and Hadwin 2005; Butler and Winne 1995; Dunlosky and Hertzog 1997; Dunlosky and Lipko 2007; Graesser et al. 2005; Metcalfe 2009; Metcalfe and Finn 2008; Nelson and Narens 1990; Schraw 2006; Veenman et al. 2006; Winne and Hadwin 1998; Zimmerman 2008).

Marie's actions have important implications for her academic success. Accurately monitoring one's own current state of learning is essential for applying effective restudy behavior and study regulation (Butler and Winne 1995; Dunlosky and Hertzog 1997; Dunlosky et al. 2005b; Nelson and Narens 1990; Powers 1973; Winne and Hadwin 1998). Learners who overestimate their current state of learning may terminate their studies prematurely, leading to poorer learning outcomes. Similarly, learners who underestimate their current state of learning may invest too much time studying information they already know while having too little time left for studying information that they have not yet learned. It is therefore important to identify and understand factors that affect how accurately learners judge their current state of learning, and how those factors influence restudy effectiveness and the recall of previously studied text information (Dunlosky et al. 2005a). The present investigation focuses on two such factors - keyword tasks and biasing titles - both of which provide cues for metacognitive monitoring and study regulation. Specifically, this investigation aims to answer the following research questions: (1) How do biasing titles and the timing of keyword tasks affect metacognitive monitoring? (2) How do biasing titles and the timing of keyword tasks affect recall improvement over two consecutive study trials? Before we engage with these topics, we discuss two ways by which students assess their current understanding of content – cue-utilization and judgments of learning.

Cue-utilization and judgments of learning

Judgments about one's own learning are referred to as judgments of learning (JoLs; Koriat 1995; Metcalfe 2002). According to the Cue-Utilization Framework (Koriat 1997; Lauterman and Ackerman 2013), learners do not monitor memory traces directly when judging the extent to which they will recall previously studied text material in the future (Kimball and Metcalfe 2003; Koriat 1995; Metcalfe 2002). Instead, learners base their judgments of learning on a variety of metacognitive cues (Koriat 1995; Metcalfe 2002). Koriat (1997) distinguishes between three types of cues: intrinsic, extrinsic, and mnemonic. *Intrinsic cues* are characteristics of the learning material that disclose the a-priori difficulty of learning. For example, the imagery value of a word is a relatively effective diagnostic of the word's memorability (Groninger 1979; Sadoski et al. 2000). It is easier to recall words that are concrete and easy to envision (such as "cat"), as opposed to words that are more abstract and difficult to envision (such as "law"). In contrast, *extrinsic cues* are factors that arise from the learning conditions, such as number of study trials, recall context (Lovelace 1984), and reading times (Mazzoni et al. 1990). For example, it is easier to predict future performance on cued recall as compared to free recall (Lovelace 1984). Finally, *mnemonic cues* indicate to learners the extent to which they will recall previously learned information in the future. Examples include the accessibility of pertinent information (Dunlosky and Nelson 1992; Koriat 1993; Morris 1990), the ease with which information comes to mind (Kelley and Lindsay 1993; Koriat 1993; Mazzoni and Nelson 1995), and the memory of previous successful recall attempts (Finn and Metcalfe 2008; Gardiner et al. 1977; King et al. 1980; Mazzoni and Cornoldi 1993). As a concrete example, having previous experience with in-class quizzes would make it easier to predict future performance on the same quiz items (Finn and Metcalfe 2008).

Intrinsic and extrinsic cues affect judgments of learning directly. Returning to Marie as she studies for her psychology exam, she might believe she will be less likely to recall information from a very long chapter (intrinsic cue), or information from chapters she has not studied as often (extrinsic cue). However, internal and external cues also exert their influence indirectly by affecting mnemonic cues (Koriat 1997). While the direct effects of intrinsic and extrinsic cues result from analytic inferences that apply to a learner's a- priori theory about the memorial consequences of a variety of factors (Koriat 1997), the effects of mnemonic cues are based on rather non-analytic inferences that employ global heuristics rather than logical, conscious deductions (Jacoby and Brooks 1984; Kelley and Jacoby 1996; Koriat 1994). For example, Marie may try to summarize a chapter based on what she remembers from it to gauge how well prepared she is for her exam. If she finds herself unable to produce a sound summary from her memory, she may decide to restudy that chapter to close her knowledge gaps (Thiede et al. 2003).

In terms of monitoring text-based learning, mnemonic cues play a particularly important role. Because learners cannot access their mental representations of entire texts when judging how much they will recall (Dunlosky et al. 2005a), the learners' judgments are likely to be influenced by access to partial information (Koriat 1995) and/or the memory of previous recall attempts (Finn and Metcalfe 2008; Gardiner et al. 1977; Mazzoni and Cornoldi 1993). Mnemonic cues are therefore the focus of this study. Specifically, we attempt to determine how learners utilize the mnemonic cues they generate when they summarize expository texts in keywords, relative to the mnemonic cues they derive from the titles of those texts.

Mnemonic cues derived from immediate vs. delayed keyword tasks

In a series of experiments on mnemonic cues, Thiede and colleagues (Thiede et al. 2003; Thiede et al. 2005) showed that learners who generate keywords after reading a number of expository texts are better able to distinguish well-learned from less well- learned texts prior to taking a first comprehension test than learners who generate keywords immediately after reading each text. Consequently, learners who generate keywords with a delay (i.e., after reading all of the texts) show greater study effectiveness in a self-regulated restudy trial and achieve higher comprehension test scores in a second test taken after the restudy trial, than learners who generate keywords immediately. This phenomenon is referred to as the Delayed-Keyword-Effect. And while Thiede et al. (2003, 2005) focused specifically on investigating the Delayed-Keyword-Effect in terms of *relative monitoring accuracy* (i.e., the extent to which learners can discriminate between well-learned and less well-learned texts; e.g., Nelson 1984), this investigation aims to determine whether the cues produced in delayed keyword tasks may also help prevent *overestimation bias* (i.e., the extent to which learners overestimate their learning and / or performance; e.g., Schraw 2009).

In line with Thiede et al. (2003), we argue that delayed keyword generation will provide more valid mnemonic cues than immediate keyword generation because the cognitive processes involved in the delayed keyword task and the first recall task are more aligned by means of a) accessing text information that is no longer highly activated in memory (Britton and

Gülgöz 1991; Fletcher et al. 1996; Van den Broek et al. 1996), and b) accessing rather consolidated mental text representations that are still accessible after the lexical and text base representations have decayed (Kintsch et al. 1990). Immediate keyword tasks, in contrast, can be performed with a highly accessible text base representation that is not indicative of performance on a delayed recall task (Thiede et al. 2005). We therefore expect that learners who generate keywords immediately - as opposed to after a delay - are more likely to overestimate their current state of learning (Table 1), and less likely to effectively regulate their restudy behavior, resulting in lower improvement from the first to the second recall task (Table 1). But keywords are not the only mnemonic cues influencing student learning. We now turn to another common mnemonic cue that learners rely on – titles.

Mnemonic cues derived from biasing titles

The mnemonic cues derived from keyword tasks are closely tied to the task conditions (i.e., the timing of the keyword tasks) and are likely to interact with mnemonic cues derived from the text material. The titles of texts, for example, function as such additional mnemonic cues (Sadoski et al. 2000). From a cognitive perspective, titles have several effects on the encoding of text information – they provide a context for upcoming text information (Ausubel 1968), activate relevant prior knowledge (Ausubel 1968), guide a reader's attention towards certain information in a text (Lorch Jr. and Lorch 1996), and provide retrieval cues for previously studied text information (Sadoski et al. 2000).

Because a title rarely highlights *all* the information that is stated in a text, titles are typically biased towards certain information in a text (Lorch Jr. 1989; Lorch Jr. and Lorch 1996; Ritchey et al. 2008). A whole body of research has investigated how biasing titles affect text encoding and retrieval (e.g., Frase and Kreitzberg n.d.; Kozminsky 1977; Lorch Jr. and Lorch 1996; Ritchey et al. 2008; Schallert 1967).

In a series of experiments, Lorch and colleagues (Lorch Jr. 1989; Lorch Jr. and Lorch 1996) demonstrated that titles specifically foster the recall of title-related information. However, title-related information was always stated at the beginning of the texts in those studies, which may have resulted in confounding effects between title-relatedness and the position of information in a text. Kieras (1978) showed that learners expect the most important information to be stated first (Initial-Mention-Effect, Kieras 1981) and information that is initially mentioned influences how learners make sense of the entire text (Kieras 1980). Ritchey et al. (2008) disentangled the effects of title-relatedness and initial-mention by generating expository texts that were comprised of two subtopics, with biasing titles highlighting either the first or the

Table 1 Hypotheses

Immediate versus Delayed Keywords

- Learners who generate keywords immediately overestimate their current state of learning to a greater extent than learners who generate keywords after a delay (particularly with regard to texts with unrelated titles).
- Learners who generate keywords immediately regulate their restudy behavior less effectively than learners who generate keywords after a delay, resulting in lower improvement from the first to the second recall task (particularly with regard to texts with unrelated titles).

Titles

- Averaged across keyword groups, learners are more likely to overestimate their recall for texts with titles that are Related/Close, followed by texts with titles that are Related/Distant, followed by texts with titles that are Unrelated.

second subtopic. The researchers found that recall for text information was facilitated when the information was related and close to titles, and inhibited when the information was unrelated to and distant from titles (Ritchey et al. 2008).

To determine whether biasing titles also affect metacognitive monitoring and study regulation, we have adopted a similar approach in this study by incorporating six expository texts with titles highlighting either the first (Related/Close) or the second subtopic (Related/Distant) of the texts. As an extension to previous works, we added a third condition in which titles are seemingly unrelated to either subtopic (Unrelated). While this condition may seem somewhat artificial at first sight, there are many real-life examples in which titles are only loosely related to their corresponding texts. An example for this is the frequent use of the German term "Erörtern" as a text-title in German textbooks (Nutz 2012). "Erörtern" describes the process of arguing for a certain position in the form of an essay. However, the meaning of the word "Erörtern" is not accessible to novices in the domain of argumentative essay writing because it offers no linguistic link to the topic it describes. More specifically, "Erörtern" is derived from the German noun 'Ort' (i.e., 'location') and has no obvious relation to essay writing. Hence, novices tend to believe that "Erörtern" is related to finding your way around rather than writing an essay according to the principles of argumentation. With respect to research on biasing titles, unrelated titles are expected to inhibit the recall of text information because they do not provide appropriate context (Ausubel 1968; Gagne 1969; Gagne and Wiegand 1970) and serve as poor retrieval cues for recall (Sadoski et al. 2000).

Interactions between biasing titles and immediate versus delayed keyword tasks

We propose that biasing titles interact with the timing of keyword tasks in how they affect the mnemonic cues learners derive from generating keywords. Because learners are more likely to generate keywords for information that is related and close to the title than unrelated or distant from it, we expect learners to be more likely to overestimate their recall for texts with titles that are Related/Close, followed by texts with titles that are Related/Distant, followed by texts with titles that are Unrelated (Table 1). Because learners who generate keywords immediately are able to draw from a highly accessible text base, we expect those learners to overestimate their recall to a greater extent than learners who generate keywords after a delay, particularly with regard to texts with unrelated titles (Table 1). It follows that learners who generate keywords immediately regulate their restudy behavior less effectively for those texts, resulting in lower improvement from the first to the second recall task for texts with unrelated titles (Table 1). In the present investigation, we investigate the role that titles and keyword tasks play in terms of metacognitive monitoring, study regulation, and recall performance.

Methods

Participants

Two hundred and thirteen undergraduate psychology students of a midsized university in the western United States participated in the study and received extra course credit. Twenty-six percent were male, 74% were female. Participants' ages ranged from 18 to 57 years (M = 22.2; SD = 3.53), and their average GPA was 3.0. Seventy-two percent of the participants were Caucasian, 20% Hispanic, 6% Asian American and 2% African American. All participants were English native-speakers.

Design and materials

The study follows a 2-between (Immediate vs. Delayed Keyword Tasks) × 3-within (Related/Close vs. Related/Distant vs. Unrelated Titles) experimental design with repeated measures on the factor "titles".

The experimental materials were composed of six expository texts derived from online databases and modified to suit the purpose of the study. Each expository text consisted of two distinct subtopics of an overall related theme. The text concerning the overall theme of navigation techniques, for example, was comprised of a subtopic on marine chronometers and a subtopic on radio navigation. Each subtopic in a text consisted of exactly 30 idea units. Idea units were defined as "single, meaningful piece[s] of information conveyed by the passage, whether [they] consisted of a word, a definition, or a phrase in the passage" (Meyer 1975). The subtopics were equated for word count (range: 190–284 words) and readability (Flesh-Kincaid readability score; range: 11–13). The readability range was chosen to match the target participant group of undergraduate university students. Each text was accompanied by one of three titles - a title that was related to the first subtopic in the text (Related/Close), a title that was related to the second subtopic in the text (Related/Distant), or a title that was unrelated to either of the subtopics in the text (Unrelated). For example, if the text on navigation techniques started with the subtopic on marine chronometers followed by the subtopic on radio navigation, then "Marine Chronometers" would be classified as a Related/Close title because it highlights the initially mentioned subtopic. "Radio Navigation" would be classified as a Related/Distant title because it highlights the second subtopic and "Lunar Tides" would be classified as an unrelated title because it does not refer to either of the subtopics stated in the text (Table 2). All titles consisted of two-noun constructions that ranged between four and seven syllables. The relatedness between a title and its corresponding subtopic was assessed in a pilot study conducted prior to the investigation and only titles that were rated as significantly more related to their corresponding subtopic than to any other subtopic were chosen for the investigation.

We made efforts to control for prior knowledge by only including topics, which are neither part of the standard US high school curriculum nor part of the standard psychology undergraduate curriculum at the university from which participants were recruited. It is possible that

Table 2 Subtopic and title variations (illustrated on the example of the text about navigation techniques)

Title Conditions	Title	Subtopic 1	Subtopic 2
Related/Close 1	Marine Chronometers	Marine Chronometers	Radio Navigation
Related/Distant 1	Radio Navigation	Marine Chronometers	Radio Navigation
Unrelated 1	Lunar Tides	Marine Chronometers	Radio Navigation
Related/Close 2	Radio Navigation	Radio Navigation	Marine Chronometers
Related/Distant 2	Marine Chronometers	Radio Navigation	Marine Chronometers
Unrelated 2	Lunar Tides	Radio Navigation	Marine Chronometers

Note: Each text cycled through two expressions of Related/Close (Related/Close 1 and Related/Close 2), Related/Distant (Related/Distant 1 and Related/Distant 2), and Unrelated (Unrelated 1 and Unrelated 2), depending on which subtopic was stated first students were exposed to the topics due to their own personal interests. However, we chose not to administer pre-tests because Campbell and Stanley (1959) indicate that the administration of pre-tests can introduce possible confounds in the form of interactions with the pre-tests and the intervention. Such confounds would be especially salient in a study of metacognition and study regulation, but can be controlled using a post-test only design with random assignment, as utilized in this study. To control for confounding effects between the type of title, the subtopic and the subtopic's position in the text, the order of title and subtopic appearance was counterbalanced within a Latin Square (Table 3). To control for confounding effects of text position, the order of text appearance was also balanced within the Latin Square.

Measures

Judgments of learning After text reading and generating keywords, participants provided a metacognitive judgment of learning that required assessing one's own current state of learning for each text on a 6-point Likert scale (1 = learned very little to 6 = learned very much). We provided only six response options because, even though Shuford and Brown (1975) recommend using a larger number of response options, providing the opportunity of transmitting more information (as offered by slide-scale percentage response scales, for example), Keren (1991) argues that in practice, assessing and comparing too many response options in a systematic manner is too difficult for participants due to limited memory and processing capacity.

Recall task 1 (recall prior to restudy) and recall task 2 (recall after restudy) Participants' recall tasks were scored for idea units using a 3-category scoring rubric. Recalled idea units were either entirely correct (1 point), partially correct (0.5 points), or incorrect / derived from prior knowledge rather than from the text (0 points). Idea units were defined as "single, meaningful pieces of information conveyed by a text, whether they consist of words, definitions, or phrases" (Meyer 1975; Ritchey et al. 2008). Each subtopic in each experimental text was constructed such that it consisted of exactly 30 distinct idea units. For example, the subtopic on radio navigation in the text on navigation techniques contained the following sentence about the OMEGA radio navigation system: "OMEGA was developed by the United

Table 3 Latin Square counterbalance design of title variations across keyword conditions

Keyword Condition												
	Immediate						Delayed					
Text	1	2	3	4	5	6	1	2	3	4	5	6
Title Variations	RC_1	RD_1	UR_1	RC_2	RD_2	UR_2	RC_1	RD_1	UR_1	RC_2	RD_2	UR_2
	RD_1	UR_1	RC_2	RD_2	UR_2	RC_1	RD_1	UR_1	RC_2	RD_2	UR_2	RC_1
	UR_1	RC_2	RD_2	UR_2	RC_1	RD_1	UR_1	RC_2	RD_2	UR_2	RC_1	RD_1
	RC_2	RD_2	UR_2	RC_1	RD_1	UR_1	RC_2	RD_2	UR_2	RC_1	RD_1	UR_1
	RD_2	UR_2	RC_1	RD_1	UR_1	RC_2	RD_2	UR_2	RC_1	RD_1	UR_1	RC_2
	UR_2	RC_1	RD_1	UR_1	RC_2	RD_2	UR_2	RC_1	RD_1	UR_1	RC_2	RD_2

RC = Related/Close; RD = Related/Distant; UR = Unrelated

Notes: Each text cycled through two expressions of RC (RC_1 and RC_2), RD (RD_1 and RD_2), and UR (UR_1 and UR_2), depending on which subtopic was stated first (see Table 2). In addition to the counterbalancing of title conditions illustrated above, the order of text appearance (1–6) was also counterbalanced in the experiment

States Navy for military aviation users." This sentence consisted of two idea units: (1) OMEGA was developed by the United States Navy; (2) Omega was developed for military aviation users. The idea units for the texts were constructed in a committee approach, which included the primary researcher and four graduate students. Because the idea units for each text and subtopic were defined prior to scoring the recall tasks, they served as the scoring guides for the two independent raters who scored the recall tasks (blind to experimental conditions).

The two independent raters who scored the recall tasks were two graduate students who had not previously been involved with this study. The raters were trained by the primary researcher on how to use the idea unit scoring guides, utilizing exemplar recall tasks from pilot studies unassociated with the to-be-scored main study data. For the data of the main study, participant IDs were first transformed into participant numbers (1–213). The recall tasks (both recall task 1 and recall task 2) were then sorted according to those participant numbers (1–213). Each of the two raters scored half of the recall tasks. Rater 1 scored the first half of recall tasks from recall task 1 (participant numbers 1–106), and the second half of recall tasks from recall task 2 (participant numbers 107–213). Rater 2 scored the second half of recall tasks from recall task 1 (participant numbers 107–213), and the first half of recall tasks from recall task 2 (participant numbers 1–106). These scores were used for the main analyses. The purpose of this scoring method was to ensure that each rater scored half of the data while eliminating biases for experimental condition (by utilizing participant numbers) and time (by having each rater score half the recall tasks of recall task 1 and the other half of recall tasks of recall task 2). To determine inter-rater reliability, rater 1 scored an additional subset of 15 recall tasks from the second half of recall task 1 (participant numbers 107–213, which had previously been scored by rater 2), and an additional subset of 15 recall tasks from the first half of recall task 2 (participant numbers 1–106, which had previously been scored by rater 2). Rater 2 scored an additional subset of 15 recall tasks from the first half of recall task 1 (participant numbers 1–106, which had previously been scored by rater 1), and an additional subset of 15 recall tasks from the second half of recall task 2 (participant numbers 107–213, which had previously been scored by rater 1). This procedure resulted in a subset of 60 recall tasks (30 from recall task 1 and 30 from recall task 2), which were redundantly scored by both raters and could be used to compute inter-rater reliability (Cohen's Kappa). Cohen's Kappa ranged from 0.74 ($p > 0.01$; 95% CI [0.833, 0.903] to 0.98 ($p < 0.01$; 95% CI [0.649, 0.821], depending upon the text on which the recall tasks were based.

The concrete scoring procedure was a follows: For each participant, the raters first segmented the recalled information from the participant's writings into idea units, using Meyer's (1975) definition as cited above. The raters then compared the idea units from the participant's writings to the scoring guide for the idea units from the corresponding experimental texts. If an idea unit from a recall task matched an idea unit from the text in its meaning, it was scored as correct (1 point). Idea units could also be scored as partially correct (0.5 points), or incorrect (0 points). For example, if a participant stated that "the OMEGA navigation system was developed by the Navy", that answer was scored as partially correct at 0.5 points, because the qualifying information that it was developed by the United States Navy (as opposed to the Royal Navy, for example), was missing. Sometimes, participants produced idea units from their own prior knowledge, which was not directly derived from information presented in the text. For example, one participant stated with regard to artwork mentioned in the text on Expressionist painting that they "saw the real piece in New York". While this information is related to the text and potentially true, it does not contribute to assessing text recall. Hence, these instances of prior knowledge were recorded, but not scored for points. The raters recorded their scoring data for each participant. The idea unit data was then re-coded for each participant, relative to that participant's counterbalance iteration of titles and subtopics (see Tables 2 and 3). For example, let us assume a participant had correctly recalled fifteen idea units from the text on navigation techniques - ten idea units from the subtopic on radio navigation and five idea units from the subtopic on marine chronometers. In that participant's set of the counterbalance design, the subtopic on radio navigation was stated first in the text (as opposed to last), and the text carried the title "Radio Navigation" (as opposed to "Marine Chronometers"). Based on this information, the ten correctly recalled idea units from the subtopic on radio navigation would be re-coded for this participant from "the number of correctly recalled idea units for radio navigation subtopic" into "the number of correctly recalled idea units that are related and close to the title". The five correctly recalled idea units from the subtopic on marine chronometers would thus be re-coded into "number of correctly recalled idea units that are unrelated to the title and distant from it". This procedure was repeated for the remaining five recall tasks, and then the mean number of correctly recalled idea units per title condition was computed.

Study procedure

The study was conducted in a university computer lab, using an experimental website and several Word documents. The procedure utilized in this study was similar to that used by Thiede et al. (2003). Upon entering the computer lab, participants were randomly assigned to either the immediate or the delayed keyword condition. Each participant read six texts and was instructed to learn as much from each of the texts as possible. Each text, along with its title, was presented for 2.5 min. After text reading, participants were prompted with only the titles again, and were asked to generate keywords. Participants were allowed to generate any number of keywords between a minimum of zero and a maximum of six keywords per text. Participants in both keywords groups generated keywords for each of the six texts. The immediate keyword group generated keywords immediately after reading each text. The delayed keyword group generated keywords only after having read all of the texts.

After reading and generating keywords, participants provided a metacognitive judgment of learning for each text, again prompted by the titles. In a next step, the text-titles were presented one at a time, and participants were asked to write about what they recalled from each text. The time limit for each of those recall-based writings was 3 min. After completing the first recall task for all six texts, participants were allotted ten minutes to go back to any of the previously studied texts for further study. Participants were able to select any number of the six texts (including zero). Participants could spend up to ten minutes restudying the texts; however, no participant met this threshold and all participants finished their restudy before the ten minutes were up. Then, the titles were presented one at a time, and participants were asked again to write about what they recalled from each text, with each recall-based writing taking no more than three minutes. Reading and writing times were controlled in order to encourage participants to engage in each task thoroughly and to prevent participants from skipping tasks. Reading and writing times were allocated according to data derived from a pilot study conducted prior to the actual investigation. For an overview of the procedure, see Table 4.

Table 4 Study procedure

Immediate Keyword Task	Delayed Keyword Task
Read text 1	Read text 1
Generate keywords for text 1	Read text 2
Read text 2	…
Generate keywords for text 2	Read text 6
…	Generate keywords for text 1
…	Generate keywords for text 2
Read text 6	…
Generate keywords for text 6	Generate keywords for text 6
Same in both experimental groups (immediate keyword group and delayed keyword group):	
Provide judgments of learning (one per text) on 6-point Likert scales	…
Recall task 1 (one recall-based writing for each of the six texts)	…
Self-regulated restudy trial	…
Recall task 2 (one recall-based writing for each of the six texts)	…

Results

Prior to performing statistical analyses we determined whether the assumptions for computing analyses of variance (ANOVAs) were sufficiently met (Bortz 1993). In some cases, the Mauchly test of sphericity indicated that the assumption of sphericity was not sufficiently met. In those cases, we corrected the reported statistical values and degrees of freedom using the Greenhouse-Geisser correction.

Metacognitive monitoring: Overestimation Bias

To compute a classical bias index (Bannert 2007, 2009; Mengelkamp and Bannert 2010; Schraw 2009; Yates 1990) we first computed each participant's mean number of correctly recalled idea units in recall task 1 for each title condition. Based on the maximum number of idea units that participants could potentially learn and recall (i.e., 60 idea units per text) we assigned each of those continuous recall performance indices to one of six recall performance categories:

Category 1: 0 ≤ number of recalled idea units <10; category 2: 10 ≤ recalled idea units <20; category 3: 20 ≤ recalled idea units <30; category 4: 30 ≤ recalled idea units <40; category 5: 40 ≤ recalled idea units <50; category 6: 50 ≤ recalled idea units ≤60. The range of category 6 encompasses one more idea unit than the other five categories. However, no participant recalled the maximum of 60 possible idea units, so the difference in range between category 6 and the other five categories was not practically relevant to data aggregation or interpretation.

We computed the bias index by subtracting a participant's mean recall performance for a title condition (interval range 1 to 6) from the mean judgment of learning that the participant provided with respect to texts in that title condition (interval range 1 to 6). We analyzed the data with a 2-keyword (Immediate vs. Delayed) × 3-title (Related/Close vs. Related/Distant vs. Unrelated) ANOVA with repeated measures on the factor "titles" and Bonferroni correction for multiple testing. The ANOVA revealed a significant main effect of keywords [F (1, 211) = 8.69; MSerror = 1.54; *p* = 0.004; partial η2 = 0.04; Fig. 1]. Participants who generated keywords immediately after reading a text overestimated their current state of learning to a greater extent (M = 2.08; SD = 0.88) than participants who generated keywords following a delay (M = 1.79; SD = 1.08), thereby supporting our hypothesis (Table 1). The ANOVA also

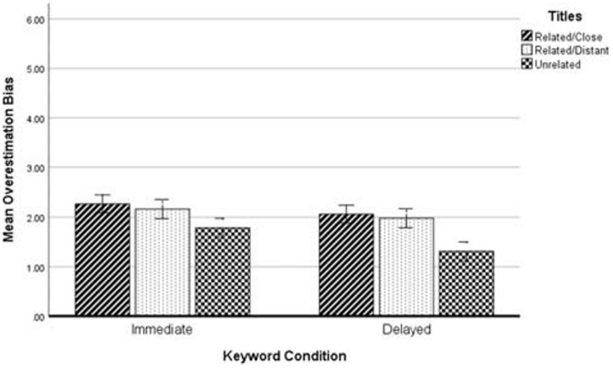


Fig. 1 Overestimation Bias at Titles (Related/Close vs. Related/Distant vs. Unrelated) by Keyword Condition (Immediate vs. Delayed)

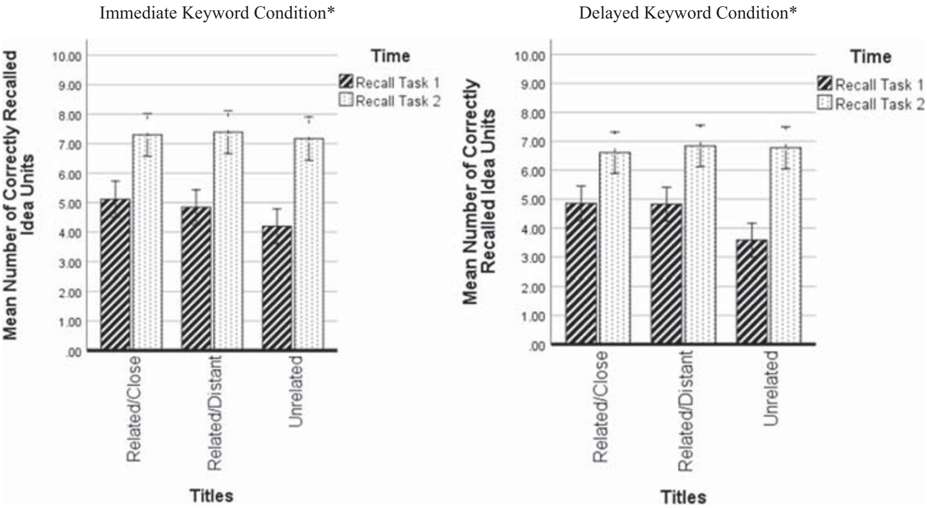
revealed a significant main effect of titles [F (1.87, 394.03) = 34.33; MSerror = 0.74; $p < 0.001$; partial η2 = 0.14; Fig. 1]. Averaged across keyword groups, participants showed a significantly stronger overestimation bias for texts with related titles (M = 2.12; SD = 0.98) than for texts with unrelated titles (M = 1.54; SD = 1.03). This result provides only partial support for our hypotheses (Table 1) because overestimation bias did not vary significantly depending on whether a title highlighted information that was stated close to it (M = 2.18; SD = 0.96) or distant from it (M = 2.09; SD = 1.00). In addition, the ANOVA failed to reveal a significant interaction between keyword conditions and titles.

For further insight into monitoring bias, we compared the number of generated keywords with a multivariate 2-keyword (Immediate vs. Delayed) × 3-title (Related/Close vs. Related/Distant vs. Unrelated) ANOVA with Bonferroni correction for multiple testing. The main effect of keywords was significant [F (1,211) = 41.73; MSerror = 1.65; $p < 0.001$; partial η2 = 0.16]. The pairwise comparisons revealed that for each of the title conditions, learners in the immediate keyword group generated significantly more keywords than learners in the delayed keyword group (all p < 0.001). Learners who generated keywords immediately generated an average of 3.96 (SD = 0.83) keywords per text and learners who generated keywords after a delay generated an average of 2.46 (SD = 1.05) keywords. A follow-up linear regression revealed that the number of keywords predicted overestimation bias [F (1, 211) = 28.23, $p < 0.001$, $R^2 = 0.12$].

Improvement from recall task 1 to recall task 2

We first computed each participant's mean number of correctly recalled idea units for each title condition in each of the two recall tasks. We analyzed the data with a 2-keyword (Immediate vs. Delayed) × 2-time (Recall Task 1 vs. Recall Task 2) × 3-title (Related/Close vs. Related/Distant vs. Unrelated) ANOVA with repeated measures on "time" and "titles" and Bonferroni correction for multiple testing. The results of the ANOVA revealed no significant main effect of keyword condition (Fig. 2), thereby failing to support our hypothesis (Table 1).

However, the results of the ANOVA revealed a significant main effect for time [F (1,211) = 69.70; MSerror = 27.36; $p < 0.001$; partial η2 = 0.25; Fig. 2]. On average, participants recalled more idea units per text in the second (final) recall task (M = 7.01; SD =3.78)


Immediate Keyword Condition*

Delayed Keyword Condition*

Fig. 2 Mean Number of Correctly Recalled Idea Units by Time (Recall Task 1 vs. Recall Task 2) and Titles (Related/Close vs. Related/Distant vs. Unrelated) at Keyword Condition (Immediate vs. Delayed). *Note: For both keyword conditions, recall improved from recall task 1 to 2. However, improvement was greater in the unrelated title condition than in the related/close and related/distant title conditions

than in the first recall task (M = 4.57; SD = 3.10), confirming that participants benefited from the restudy trial to increase their recall performance. The results also revealed a significant main effect of titles [F (1.80, 380.11) = 7.99; MSerror = 5.74; p < 0.001; partial η2 = 0.36]. Averaged across recall tasks, participants recalled more idea units from texts with related (M = 5.98; SD = 3.44), than from texts with unrelated titles (M = 5.43; SD = 3.45).

These two main effects were further qualified by a significant interaction between titles and time [F (2, 421.91) = 10.40; MSerror = 3.40; *p* < 0.001; partial η2 = 0.05; Fig. 2]. Participants achieved higher levels of recall improvement for texts with unrelated titles (Mdiff = 3.08; SD = 3.45), as compared to texts with titles that highlight information stated close to them (Mdiff =1.97; SD = 3.46), or distant from them (Mdiff = 2.28; SD = 3.41). Said another way – while the unrelated titles initially led to the poorest recall, when given a chance to re-study these materials, students were able to increase their recall, on average, to a level comparable to those who saw titles related to the text content (Fig. 2).

To gain more insight into subtopic-specific recall improvement, we conducted two sets of additional analyses. First, we computed a 2-keyword (Immediate vs. Delayed) × 2-time (Recall Task 1 vs. Recall Task 2) × 2-title (Related vs. Unrelated) × 2-subtopic (Close vs. Distant) ANOVA with repeated measures on "time", "title", "subtopic", and Bonferroni correction for multiple testing. Note that with regard to subtopics, "close" refers to subtopics that were initially mentioned (i.e., close to the title), regardless of whether they were related to the title or not (Table 2). "Distant" refers to subtopics that appeared in second position in a text (i.e., distant from the title), regardless of whether they were related to the title or not (Table 2). In addition to confirming the above described effects of time and titles, this ANOVA revealed a significant main effect of subtopic [F (1, 211) = 157.20; MSerror = 5.59; *p* < 0.001; partial η2 = 0.43]. Averaged across recall tasks, participants recalled more idea units from subtopics that were close to titles (M = 3.63; SD = 2.27) than from subtopics that were distant from titles (M = 2.34; SD = 2.19), thus indicating that initial-mention effects were operating in addition to title effects. The ANOVA also revealed a significant two-way interaction between keyword conditions and subtopics [F (1, 211) = 9.24; MSerror = 4.59; $p = 0.003$; partial η2 = 0.04], which was further qualified by a three-way interaction between keywords, time, and subtopics [F (1, 211) = 16.75; MSerror = 4.59; $p < 0.001$; partial η2 = 0.07].

Separately for the immediate and delayed keyword conditions, we followed up with 2-time (Recall Task 1 vs. Recall Task 2) × 2-subtopic (Close vs. Distant) ANOVAs with repeated measures on “time” and “subtopic” and Bonferroni corrections for multiple testing. For the immediate keyword condition, the ANOVA revealed a main effect of time [F (1, 104) = 29.68; MSerror = 19.78; p < 0.001; partial η2 = 0.22; Fig. 3], indicating that, averaged across subtopics, recall improved from recall task 1 (M = 4.98; SD = 3.38) to recall task 2 (M = 7.34; SD = 4.06). The ANOVA also revealed a main effect of subtopics [F (1, 104) = 53.63; MSerror = 7.61; p < 0.001; partial η2 = 0.34; Fig. 3]. Averaged across time, recall was higher for subtopics that were close (M = 7.15; SD = 4.21) as compared to distant from a title (M = 5.18; SD = 3.23). These two main effects were qualified by a significant interaction between time and subtopics [F (1, 104) = 8.95; MSerror = 11.07; p = 0.003; partial η2 = 0.08; Fig. 3]. In the immediate keyword condition, recall of information that was distant from the title improved from recall task 1 (M = 3.51; SD = 2.81) to recall task 2 (M = 6.84; SD = 3.65) and became similar to the recall of information that was close to the title in recall task 2 (M = 7.84; SD = 4.47). This indicates that learners in the immediate keyword group closed the gap between recall of initially mentioned information and information stated later in a text in the restudy trial (Fig. 3).

For the delayed keyword condition, the ANOVA also revealed main effects of time [F (1, 107) = 21.97; MSerror = 17.43; p < 0.001; partial η2 = 0.17; Fig. 3] and subtopics [F(1, 107) = 105.56; MSerror = 10.69; p < 0.001; partial η2 = 0.50; Fig. 3], in the same directions as

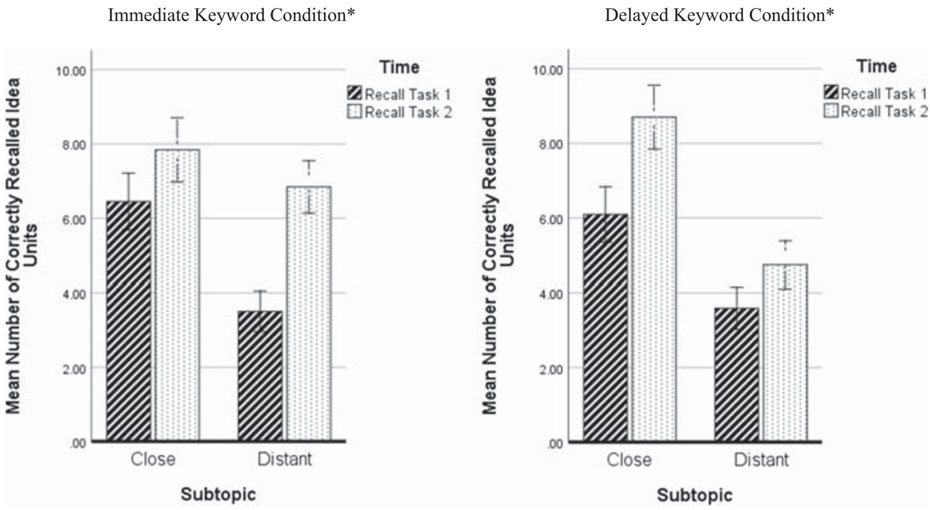


Fig. 3 Mean Number of Correctly Recalled Idea Units by Keyword Condition (Immediate vs. Delayed) and Time (Recall Task 1 vs. Recall Task 2) and Subtopic (Close vs. Distant). *Note: Recall for both types of subtopics (close and distant) improved from recall task 1 to 2. However, in the immediate keyword condition, recall for distant subtopics improved more than recall for close subtopics, so that recall for close and distant subtopics was similar in recall task 2. In the delayed keyword condition, recall for close subtopics improved more than recall for distant subtopics, thus retaining the pattern of greater recall for close over distant subtopics from recall task 1 to 2

observed in the immediate keyword group (Fig. 3). The time-by-subtopic interaction was also significant [F (1, 107) = 7.80; MSerror = 7.34; $p = 0.006$; partial η2 = 0.07; Fig. 3], but reversed in direction when compared to the immediate keyword group. In contrast to learners in the immediate keyword group, learners in the delayed keyword group did not use the restudy time to close the gap between the recall of initially mentioned information and the recall of information stated later in the text. Instead, they further increased their recall for initially mentioned information (Fig. 3).

Discussion

Following the assumptions in the Cue-Utilization-Framework (Koriat 1997; Lauterman and Ackerman 2013), we expected that learners would utilize the mnemonic cues (Koriat 1997) derived from the keyword tasks in combination with biasing titles to inform their judgments of learning, which would presumably affect overestimation bias (Gardiner et al. 1977; King et al. 1980; Mazzoni and Cornoldi 1993) and improvement over recall tasks (Dunlosky and Hertzog 1997; Thiede et al. 2003, 2005).

Overestimation Bias: Discussion, conclusions and limitations

Based on prior research (e.g., Kintsch et al. 1990; Thiede et al. 2003, 2005) we expected learners in the immediate keyword group to exhibit a stronger overestimation bias than learners in the delayed keyword group (Table 1), and the results of this study support this main hypothesis. This finding extends current research by demonstrating that delayed keyword tasks do not only aid learners in distinguishing well-learned from less well-learned texts (Thiede et al. 2003, 2005) but also help prevent overestimation bias. We propose that delayed keyword tasks help prevent overestimation bias because to generate delayed keywords, learners have to access consolidated mental text representations that are available after the lexical and text base representations have decayed (Kintsch et al. 1990; Thiede et al. 2003, 2005). Immediate keyword tasks, in contrast, can be performed with highly accessible lexical and text base representations, which decay over time, and therefore serve as poor indicators of future recall performance (Kintsch et al. 1990; Thiede et al. 2003, 2005) – a theoretical notion empirically underlined by the results of this study. More nuanced follow-up analyses, which took into account the number of keywords learners generated, further confirmed that learners based their judgments of learning on cues they generated in the keyword task. Learners in the immediate keyword group generated more keywords than learners in the delayed keyword group and the number of keywords learners generated predicted overestimation bias.

In addition to this main effect of keywords, we expected a main effect of biasing titles on overestimation bias. Based on cognitive research on biasing titles (Lorch Jr. and Lorch 1996; Lorch Jr. 1989; Ritchey et al. 2008), we specifically hypothesized that learners would be more likely to overestimate their recall for texts with titles that are Related/Close, followed by texts with titles that are Related/Distant, followed by text with unrelated titles (Table 1). This hypothesis was only partially supported. Overestimation bias was significantly lower for texts with unrelated as compared to related titles but, contrary to our expectations, overestimation bias did not vary significantly depending on whether title-related information was stated close or distant in the text. We interpret this finding in the context of possible primacy/recency effects (Bjork and Whitten 1974; Brodie and Murdock 1977; Howard and Kahana 1999).

Learners tend to recall information stated first and last in a text more readily than information that is stated in the middle. The Related/Close and Related/Distant titles may have produced more similar mnemonic cues than we had previously anticipated. In the future, we plan to extend this line of research with titles targeting the close, middle, and distant segments of texts.

A limitation to the interpretation of our results on overestimation bias pertains to the monitoring accuracy measure used in this study. This study focused on absolute monitoring accuracy (Schraw 1995). Measures of absolute monitoring accuracy refer to the difference between learners' judgments of learning and recall performance (Mengelkamp and Bannert 2010). They can be computed using several methods (e.g., absolute accuracy index, Schraw 2009; bias index, Schraw 2009; Hamann coefficient, Nietfeld et al. 2006), all of which yield strengths and limitations (Mengelkamp and Bannert 2010). Because we were specifically interested in overestimation bias, we focused on the classical bias index (Schraw 2009; Yates 1990) because it allows for computing the *signed* difference between judgments of learning and actual test performance. This study focused on investigating judgments of learning that pertain to entire texts rather than specific items in the text. This required learners to provide global instead of local judgments (for an overview of global vs. local judgments of learning see Dunlosky et al. 2005a; for a recent discussion on how granularity matters in measuring self-regulated learning, see Rovers et al. 2019). We provided the learners with six rather broad response options for their judgments (1 = learned very little to 6 = learned very much). We compared those judgments to learners' actual recall performance, which was mapped onto the same six categories, as often performed in standards-based grading (Tomlinson and McTighe 2006). One could argue that we might have obtained a more sensitive overestimation measure if we had provided learners with more than six categories. However, our decision was based on Keren's work (Keren 1991), suggesting that providing learners with too many response options impedes the learners' ability to compare the response options in a systematic manner due to limited memory and processing capacity.

Improvement over recall tasks: Discussion, conclusions and limitations

Discrepancy Reduction Models of Self-Regulated Learning (e.g., Dunlosky and Hertzog 1997; Nelson and Narens 1990) propose that learners who monitor their learning more accurately also regulate their learning activities in a restudy trial more effectively and show higher levels of improvement and superior final comprehension test performance than learners who monitor their learning less accurately. In line with the findings of Thiede et al. (2003, 2005), the results of this study revealed that learners who generated keywords with a delay were able to monitor their learning more accurately than learners who generated keywords immediately. However, in terms of improvement from recall task 1 to recall task 2 (as measured by the mean difference between correctly recalled idea units per text), we detected no differences between learners in both keyword groups. Contrary to our expectations, the lower overestimation bias observed in the delayed keyword group did not result in higher levels of overall recall improvement and higher final overall recall performance as compared to the immediate keyword group. However, this result may be at least partially due to the experimental design. Learners in *both* keyword groups were encouraged to participate in a restudy trial. It was possible for participants to select zero texts for restudy (i.e., skip the restudy trial), but the experimental procedure likely facilitated restudy behavior by prompting participants that they were "now entering the restudy portion of the study", and presenting links to the texts, which made it easy to select texts for restudy. In addition, the study took part in a lab setting, which may have encouraged socially desirable restudy behavior in our participants, who were students. If the experimental setting had been less encouraging to participate in the restudy trial, it is possible that learners who generated keywords immediately and exhibited a stronger overestimation bias would have shown a greater tendency to skip restudy and show lower levels of improvement and final recall performance than learners who generated keywords with a delay (e.g., Dunlosky and Lipko 2007). We plan to examine this in a subsequent study.

Additionally, it is important to take a closer look at the differential effects of the timing of the keyword task on the number of correctly recalled idea units *per subtopic*. Prior to the self-regulated restudy trial, for texts with Related/Close and Related/Distant titles, learners in both keyword groups recalled the largest number of idea units from subtopics that were stated close to the title, whether those subtopics were related to the title or not. Learners in both keyword groups were able to improve their recall performance from recall task 1 to recall task 2, but only learners in the immediate keyword group showed higher levels of improvement for subtopics that were stated distant from the title. Consequently, only the learners in the immediate keyword group seem to have used the restudy trial in order to gain knowledge on *all* the subtopics in a text, rather than solely focusing on initially mentioned information (Kieras 1978, 1980, 1981). For learners in the delayed keyword group, the initial-mention-effect (Kieras 1978) was observed even after the restudy trial by showing higher levels of improvement for subtopics that were stated close to the title, regardless of title-relatedness. This pattern of results suggest that learners applied different restudy strategies, depending on whether they generated keywords immediately after reading, or with a delay. A possible explanation for this effect can be found in the research of Jacoby and Bartz (1972) and Thiede et al. (2011). Thiede et al. (2011) demonstrated that test performance varies with the nature of the test that learners expect. Jacoby and Bartz (1972) propose that learners process verbal information differently, depending on whether they expect to be tested immediately after reading, or after a delay. The results of this study provide some indication that learners who generated keywords immediately after reading may have deliberately chosen to focus on initially mentioned text information in the first study trial because focusing on initially mentioned information (and disregarding information stated later) would have provided them with extended rehearsal time (Baddeley and Logie 1999) and higher chances to successfully generate a large number of keywords after reading. Following this notion, learners in the immediate keyword group may have generated additional metacognitive cues regarding their strategy to encode information in the first study trial by deliberately focusing on initially mentioned information. Being metacognitively aware of applying this particular study strategy may have encouraged the learners to reverse their study strategy by focusing on information stated late in the text in the restudy trial in order to improve their overall text recall.

In contrast to learners in the immediate keyword group, learners who generated keywords with a delay may have focused on initially mentioned information in a less deliberate manner. For example, Glanzer et al. (1984) showed that learners generally spend more time reading the first paragraph of a text than reading the following paragraphs. Kieras (1978) demonstrated that learners typically read the first sentences of a text slower than the following sentences regardless of whether those first text sentences are related to the text topic or not. Being metacognitively unaware of focusing on initially mentioned information may have prevented the learners in the delayed keyword group from adapting their study strategy in the restudy trial (e.g., Winne and Hadwin 1998) and may have therefore fostered the perseverance of the initial-mention-effect (Kieras 1978). However, because we were not able to collect data on the learners' actual restudy behavior, we cannot conclude whether the different restudy strategies observed in the two keyword groups were deliberately chosen and applied by the learners, or whether the use of those strategies emerged less deliberately from the interaction with the learning task. More research is needed to clarify the extent to which immediate versus delayed keyword tasks and biasing text-titles promote the deliberate use of different study and restudy strategies.

Implications

The results of this study highlight the importance of choosing keyword strategies with respect to educational goals. Learners overestimate their current state of learning to a greater extent when they generate keywords immediately after reading a text. If the educational goal is to facilitate monitoring accuracy, educators should encourage learners to generate keywords with a delay.

Regarding the design of educational text materials, the results of this study highlight the importance of mentioning information that is most important at the beginning of an expository text because learners are more likely to recall initially mentioned information. If an expository text (as found in textbooks or web resources like Wikipedia) consists of multiple subtopics (e.g., marine chronometers and radio navigation) pertaining to one overarching theme (e.g., navigation techniques), learners are more likely to recall information from the subtopic that is stated first, whether that subtopic is highlighted by the text-title or not.

The study further indicates that different keyword strategies encourage learners to apply study and restudy behavior that affects the subtopic-specific recall of information from the text. Immediate keyword tasks encourage learners to specifically focus on initially mentioned information in a first study trial, and reverse that study strategy in a restudy trial. Delayed keyword tasks encourage learners to continuously focus on initially mentioned text information during the study as well as the restudy phase. If keyword tasks are implemented in educational settings, the timing of the keyword tasks should be chosen carefully, depending on whether a learning task requires recalling primarily information from the first part of a text, or whether the learning goal is to remember information that is spread across the entire text.

Finally, the results of this study show that learners monitored their learning most accurately and showed the highest level of improvement for texts with titles that were unrelated to the topics discussed in the text. While it may not be recommendable to provide learners with texts with unrelated titles in real-life learning settings, it is still interesting to consider the theoretical implications of this finding. Highly abstract titles, for example, may function similarly to unrelated titles in that they might not aid learners in establishing an immediate connection between the title and the text contents and may therefore impair the encoding and retrieval of information stated in the text (Lippmann et al., 2019). Following this notion it might be interesting to consider whether there are features of text-titles (such as the degree of title abstractness) that might foster monitoring accuracy by means of decreasing the ease-of-processing while providing encoding and retrieval benefits (such as encouraging deeper-level-processing; Kintsch et al. 1990) instead of inhibiting encoding and retrieval the way that unrelated titles do.

Funding This research was funded by the German Academic Exchange Service / Deutscher Akademischer Austauschdienst (DAAD), grant number: D/09/45081.

Compliance with ethical standards

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

Research involving human participants This research was conducted in accordance with the ethical standards of the university and country at which this study was conducted. Prior to conducting this research, approval of the university's Institutional Review Board (IRB) was obtained. Participants volunteered for this research, with the option of extra course credit.

Informed consent and debriefing Participants provided their informed consent prior to participating in this study. Immediately after participating, participants received a detailed debriefing about this study.

References

Ausubel, D. (1968). *Educational psychology: A cognitive view*. New York: Holt, Rinehart, and Winston.

Azevedo, R. (2009). Theoretical, conceptual, methodological, and instructional issues in research on metacognition and self-regulated learning: A discussion. *Metacognition and Learning, 4*(1), 87–95.

Azevedo, R., & Hadwin, A. F. (2005). Scaffolding self-regulated learning and metacognition: Implicationsfor the design of computer-based scaffolds. *Instructional Science, 33*, 367–379.

Baddeley, A. D., & Logie, R. H. (1999). Working memory: The multiple-component model. In: A. Miyake, P. Shah (Eds.), *Models of Working Memory*, (pp. 28–61). Cambridge University Press.

Bannert, M. (2007). *Metakognition beim Lernen mit Hypermedia. Erfassung, Beschreibung und Vermittlung wirksamer metakognitiver Lernstrategien und Regulationsaktivitäten [Metacognition and learning with hypermedia]*. Münster: Waxmann.

Bannert, M. (2009). Promoting self-regulated learning through prompts: A discussion. *Zeitschrift für Pädagogische Psychologie, 23*, 139–145.

Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology, 6*(2), 173–189. https://doi.org/10.1016/0010-0285(74)90009-7.

Bortz, J. (1993). *Statistik für Sozialwissenschaftler (statistics for social Scientistis)*. Berlin: Springer.

Britton, B. K., & Gülgöz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology, 83*, 329–345.

Brodie, D. A., & Murdock, B. B. (1977). Effects of presentation time on nominal and functional serial-position curves in free recall. *Journal of Verbal Learning and Verbal Behavior, 16*(2), 185–200. https://doi.org/10.1016/s0022-5371(77)80046-7.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*, 245–281.

Campbell, D., & Stanley, J. (1959). *Experimental and quasi-experimental design for research*. Skokie: Rand McNally.

Dunlosky, J., & Hertzog, C. (1997). Older and younger adults use a functionally identical algorithm to select items for restudy during multitrial learning. *Journal of Gerontology: Psychological Science, 52*, 178–186.

Dunlosky, J., Hertzog, C., Kennedy, M., & Thiede, K. (2005a). The self-monitoring approach for effective learning. *Cognitive Technology, 10*, 4–11.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*(4), 228–232.

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005b). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language, 52*, 551–565.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOLs) and the delayed-JOL effect. *Memory & Cognition, 20*, 373–380.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58*, 19–34.

Fletcher, C. R., van den Broek, P., & Arthur, E. J. (1996). A model of narrative comprehension and recall. In B. K. Britton & A. C. Graesser (Eds.), *Models of understanding text* (pp. 141–164). Mahwah: Erlbaum.

Frase, L. T., & Kreitzberg, V. S. Effect of topical and indirect learning directions on prose recall. *Journal of Educational Psychology, 67*, 320–324.

Gagne, R. M. (1969). Context, isolation and interference effects on the retention of prose. *Journal of Educational Psychology, 60*, 408–414.

Gagne, R.M. & Wiegand, W.K. (1970) The effect of superordinate contexts on learning and retention of facts. *Journal of Educational Psychology, 61*, 406–409.

Gardiner, J. M., Passmore, C., Herriot, P., & Klee, H. (1977). Memory for remembered events: Effects of response mode and response-produced feedback. *Journal of Verbal Learning and Verbal Behavior, 16*, 45–54.

Glanzer, M., Fischer, B., & Dorfman, D. (1984). Short-term storage in reading. *Journal of Verbal Learning and Verbal Behavior, 23*, 467–486.

Graesser, A., McNamara, D., & VanLehn, K. (2005). Scaffolding deep comprehension strategies throughPoint and Query, AutoTutor and iSTART. *Educational Psychologist, 40*, 225–234.

Groninger, L. D. (1979). Predicting recall: The 'feeling-that-I-will-know' phenomenon. *American Journal of Psychology, 92*, 45–58.

Howard, M. W., & Kahana, M. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition., 25*(4), 923–941. https://doi.org/10.1037/0278-7393.25.4.923.

Jacoby, L. L., & Bartz, W. H. (1972). Rehearsal and transfer to LTM. *Journal of Verbal Learning and Verbal Behavior, 11*, 561–565.

Jacoby, L. L., & Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception and concept learning. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 18, pp. 1–47). New York: Academic Press.

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory and Language, 35*, 157–175.

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32*, 1–24.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77*, 217–273.

Kimball, D. R., & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory & Cognition, 31*, 918–929.

Kieras, D. E. (1978). Good and bad structure in simple paragraphs: Effects on apparent theme, reading time, and recall. *Journal of Verbal Learning and Verbal Behavior, 17*, 13–28.

Kieras, D. E. (1980). Initial mention as a signal to thematic content in technical passages. *Memory and Cognition, 8*, 345–353.

Kieras, D. E. (1981). Topicalization effects in cued recall of technical prose. *Memory and Cognition, 9*, 541–549.

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *American Journal of Psychology, 93*, 329–343.

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language, 29*, 133–159.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*, 609–639.

Koriat, A. (1994). Memory's knowledge of its own knowledge: The accessibility account of the feeling of knowing. In J. Metcalfe & P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 115–135). Cambridge: MIT Press.

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General, 124*, 311–333.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*, 349–370.

Kozminsky, E. (1977). Altering comprehension: The effect of biasing titles on text comprehension. *Memory & Cognition, 5*(4), 482–490.

Lauterman, T. & Ackerman, R. (2013). *Overcoming screen inferiority in text learning*. In Knauff, M., Pauen, N., Sebanz, & I. Wachsmuth (Eds.) Proceedings of the 35th Annual Conference of the Cognitive Science Society (p. 2914–2919). Austin TX: Cognitive Science Society.

Lippmann, M., Schwartz, N. H., Jacobson, N. G., Narciss, S. (2019). The concreteness of titles affects metacognition and study motivation. *Instructional Science, 47*(3), 257–277.

Lorch Jr., R. F. (1989). Text-signaling devises and their effects on reading and memory processes. *Educational Psychology Review, 1*, 209–234.

Lorch Jr., R. F., & Lorch, E. P. (1996). Effects of organizational signals on free recall of expository text. *Journal of Educational Psychology, 88*, 38–48.

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 756–766.

Mazzoni, G., & Cornoldi, C. (1993). Strategies in study time allocation: Why is study time sometimes not effective? *Journal of Experimental Psychology: General, 122*, 47–60.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory & Cognition, 18*, 196–204.

Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1263–1274.

Mengelkamp, C., & Bannert, M. (2010). Accuracy of confidence judgments: Stability and generality in the learning process and predictive validity for learning outcome. *Memory & Cognition, 38*(4), 441–451.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131*, 349–363.

Metcalfe, J. (2009). Metacognitive Judgments and Control of Study. Current *Directions in Psychological Science, 18*(3), 159–163.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*(1), 174–179.

Meyer, B. J. F. (1975). *The organization of prose and its effects on memory*. New York: American Elsevier Publishing.

Morris, C. C. (1990). Retrieval processes underlying confidence in comprehension judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 223–232.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109–133.

Nietfield, J. L., Enders, C. K., & Schraw, G. (2006). A Monte Carlo Comparison of Measures of Relative and Absolute Monitoring Accuracy. *Educational and Psychological Measurement, 66*(2).

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 26, pp. 125–173). San Diego: Academic Press.

Nutz, M. (2012). deutsch.werk.4 Sprach- und Lesebuch (German text book for students in grade 8). Leipzig: Ernst Klett Schulbuchverlag Leipzig GmbH.

Powers, W. T. (1973). *Behavior: The control of perception*. Chicago: Aldine.

Ritchey, K., Schuster, J., & Allen, J. (2008). How the relationship between text and headings influences readers‘memory. *Contemporary Educational Psychology, 33*, 859–874.

Rovers, S. F., Clarebout, G., Savelberg, H. H., de Bruin, A. B., & van Merriënboer, J. J. (2019). Granularity matters: Comparing different ways of measuring self-regulated learning. *Metacognition and Learning, 14*, 1–19.

Sadoski, M., Goetz, E. T., & Rodriguez, M. (2000). Engaging texts: Effects of concreteness on comprehensibility, interest, and recall in four text types. *Journal of Educational Psychology, 92*(1), 85–95.

Schallert, D. L. (1967). Improving memory for prose: The relationship between depth of processing and context. *Journal of Verbal Learning and Verbal Behavior, 15*, 621–632.

Schraw, G. (1995). Measures of Feeling-of-Knowing accuracy: A new look at an old problem. *Applied Cognitive Psychology, 9*, 321–322.

Schraw, G. (2006). *Knowledge: Structures and processes*. In: Alexander, P. A., Winne, P.H. (Eds.) Handbook of Educational Psychology (pp. 245–263). New York: Routledge.

Schraw, G. (2009). Measuring metacognitive judgments. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of Metacognition in Education* (pp. 415–429). New York, NY: Routledge.

Shuford, E., & Brown, T. A. (1975). Elicitation of personal probabilities and their assessment. *Instructional Science, 4*(2), 137–188.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66–73.

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the Delayed-Keyword-Effect on metacomprehension accuracy. *Journal of Experimental Psychology, Learning, Memory, and Cognition, 31*(6).

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology, 82*, 264–273.

Tomlinson, C., & McTighe, J. (2006). *Integrating differentiated instruction and understanding by design*. Alexandria: ASCD.

Van den Broek, P., Risden, K., Fletcher, C. R., & Thurlow, R. (1996). A “landscape” view of reading: Fluctuating patterns of activation and the construction of a stable memory representation. In B. K. Britton & A. C. Graesser (Eds.), *Models of understanding text* (pp. 165–187). Mahwah: Erlbaum.

Veenman, M. V. J., Van Hout-wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning*, *1*,3–1,114.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice. The educational psychology series* (pp. 277–304). Mahwah: Lawrence Erlbaum Associates Publishers.

Yates, J. F. (1990). Judgment and decision making. Prentice-Hall, Inc.

Zimmerman, B. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Journal of International Research, 45*, 166–183.

Affiliations

Marie Lippmann[1] • **Robert W. Danielson**[2] • **Neil H. Schwartz**[1] • **Hermann Körndle**[3] • **Susanne Narciss**[3]

Robert W. Danielson
robert.danielson@wsu.edu

Neil H. Schwartz
nschwartz@csuchico.edu

Hermann Körndle
Hermann.Koerndle@tu-dresden.de

Susanne Narciss
Susanne.Narciss@tu-dresden.de

[1] Department of Psychology, California State University, Chico, 400 West First Street, Chico, CA 95929, USA

[2] College of Education, Washington State University, 412 E Spokane Falls Blvd, Spokane, WA 99202, USA

[3] Department of the Psychology of Learning and Instruction, Technische Universität Dresden, Zellescher Weg 17, 01062 Dresden, Germany