Instructional Science (2025) 53:337–364
https://doi.org/10.1007/s11251-024-09686-4

ORIGINAL RESEARCH

Comparing the effectiveness of multiple text reading and rereading on knowledge retention and metacognitive accuracy

Peter Seban[1,3] · **Kamila Urban**[2] · **Radovan Sikl**[3]

Received: 9 January 2023 / Accepted: 19 September 2024 / Published online: 12 November 2024

Abstract

Although learning approaches are designed to enhance individuals' ability to store and retrieve information, not all of them are considered effective. The goal of the present study was to experimentally compare the test performance as well as the accuracy of metacognitive judgements of a multiple text reading group, rereading group, and single reading group in a one-day vs. one-week delayed test (3 × 2 between subject design). A total of 186 psychology students (M_{age} = 20.76) participated in the experiment focused on reading comprehension, accuracy of metacognitive judgments and knowledge retention. Results of the knowledge test indicate that in the one-day delayed test, multiple text reading and rereading yield similar results: both deliver slightly better results than single reading. In the one-week delayed test, though, multiple text reading yields better results than rereading, and both these reading approaches outperform single reading. Moreover, multiple text reading results in fairly robust knowledge retention with only a slight decrease in scores between the one-day delayed and one-week delayed test. Regarding metacognitive monitoring, judgements of learning in the multiple text reading group remained relatively stable after each reading and participants were underconfident about their knowledge. In the rereading group, judgements of learning increased after each reading and participants were overconfident about their knowledge, especially on the one-day delayed test. These findings have implications for educational practices aimed at enhancing learning outcomes and promoting effective learning.

Keywords Self-regulated learning · Effective learning · Multiple text reading · Rereading · Judgements of learning · Distributed practice

✉ Peter Seban
seban1@ucm.sk

[1] Department of Psychology, University of Ss. Cyril and Methodius in Trnava, Namestie J. Herdu 2, 917 01 Trnava, Slovakia

[2] Institute for Research in Social Communication, Slovak Academy of Sciences, Bratislava, Slovakia

[3] Institute of Psychology, Czech Academy of Sciences, Prague, Czechia

Introduction

Learning is a tool which humans and other organisms use to adapt to and cope with their environment. The complex society in which we, humans, live in forces us to constantly acquire and update our knowledge. In this sense, formal schooling, which is limited mainly to the early stages of life, quickly becomes outdated. The learning process, however, continues throughout life and becomes mostly self-regulated (Pressley & Woloshyn, 1995). Self-regulated learning (SRL) is characterized by setting of own study goals, managing own time, making decisions about the strategies and techniques to be used, and by monitoring and regulating of learning to attain the self-defined goals (Schraw et al., 2006; Winne, 2015; Zimmermann, 1990). While SRL broadly encompasses motivational, cognitive, and metacognitive aspects of learning (see Panadero, 2017; Schraw et al., 2006), the present study focuses on the cognitive aspect of SRL, mainly multiple text reading and rereading, along with their impact on metacognitive judgments (judgments of learning; JOLs). SRL plays a crucial role in effectively managing the reading of multiple texts. It requires proactive and deliberate planning, monitoring, and regulation of one's cognitive processes and behaviors to meet learning objectives. Employing SRL strategies can greatly improve students' abilities to understand and synthesize information from various texts (Lee & List, 2023). The goal of the present study is to investigate the short and long-term effectiveness of two approaches to reading, namely multiple text reading and rereading, among separate groups of participants. Our focus centers on evaluating the participants' performance on a knowledge test, with a particular emphasis on the depth of comprehension, while simultaneously examining the accuracy of their metacognitive monitoring. This between-subjects design will allow us to compare how these two reading approaches impacted participants' knowledge retention and metacognitive judgments differently across a short-term (occurring within a single day) and a more protracted long-term (spanning a week) timeframe.

We aim to explore the effectiveness of multiple text reading within a controlled setting. Our primary objective is to assess its potential in enhancing knowledge retention. Essentially, we are replicating a common real-world situation wherein individuals are faced with the choice of either revisiting a previously read text or seeking out an alternative source that discusses the same topic. This practice, referred to as "multiple documents" or multiple text reading (Castells et al., 2021; Richter & Maier, 2017), involves the use of texts that describe one situation from different or similar perspectives, offering unique variations in sentence structure and wording while addressing the same topic.

Rereading

One widely used learning approach is rereading (Bjørk et al., 2013; Dunlosky et al., 2013), which is preferred and commonly utilized by students (Carrier, 2003; Hartwig & Dunlosky, 2012; Karpicke et al., 2009). There are numerous reasons why rereading is so favored by the learners. Firstly, this technique is completely intuitive and requires no prior training (Dunlosky et al., 2013). Secondly, once the learner is acquainted with the text, rereading takes less time than the initial reading, which is why the technique seems time-effective (Dunlosky et al., 2013). Thirdly, reviewing the same material multiple times produces a subjective feeling of knowledge acquisition (Bjørk et al., 2013; Koriat, 1997; Wiley et al., 2005).

Studies on the effectiveness of rereading produced mixed results. Some tried to find out whether rereading is more effective than a single reading. For instance, Rawson and Kintsch (2005) found that massed rereading, i.e., rereading immediately after first reading, has benefits over a single reading when it is immediately followed by a test, while distributed rereading (with one week between readings) provides benefits when the test is delayed by two days. On the other hand, Callender and McDaniel (2009) found no beneficial effect of two rereadings on immediate performance (a test and a summary). Regarding the quality of representation gained by rereading, Fritz et al. (2000) demonstrated that rereading a text three times (each reading a week apart) had no effect on extraction of further knowledge and led to no change in the initial representation of the text. Regarding the distribution (spacing) versus massing of rereading, Glover and Corkill (1987) and Krug et al. (1990) showed that distributed rereading had a greater positive impact on learners' performance in an immediate free-recall test than massed rereading did.

Despite its popularity, rereading has some drawbacks due to which it is considered a low-utility technique. Most importantly, in a direct comparison with active learning techniques (such as elaborative interrogation, practice testing, or self-explanation) rereading led to worse results in the long-term perspective (e.g., Bjørk et al., 2013; Butler, 2010; Dunlosky et al., 2013; McDaniel et al., 2012; Weinstein et al., 2010). According to Wiley et al. (2005), rereading is the main source of unrealistic judgements about one's own performance. By seeing a text repeatedly, learners feel that they had already mastered the material and do not pay as much attention as they did upon their first encounter. Specifically, Wiley et al. (2005) noted that during the second reading, learners experience a greater perceptual fluency, which is misjudged for understanding. In short, students tend to mindlessly wander over a well-known text only to reassure themselves that their knowledge of the subject is sufficient. The overall result is a false sense of understanding the subject in question, which is why learners feel no need study any further. Consequently, they often terminate the learning without acquiring durable knowledge.

Multiple text reading

The use of multiple textual sources used to be the domain of experts, especially historians. Nowadays, though, the availability of a great number of learning resources (textbooks or internet educational pages) made the use of multiple texts more common for ordinary students and, in fact, curricula call for improvements in students' ability to work this way (OECD, 2016). Multiple text reading is an approach which considers many aspects of working with various textual resources. Bråten et al. (2020, p. 80) define it as follows: "Reading multiple texts involves trying to construct meaning from multiple textual resources that present consistent, componential (i.e., information across different texts is part of a larger whole not specified in any single text), or conflicting information on the same situation, issue, or phenomenon". Britt and Rouet (2020, p. 1) further contribute to our understanding of multiple text reading with their definition: "Multiple document comprehension refers to people's acquisition of information from more than one document for the purpose of achieving their goals. In multiple document comprehension, documents are associated with distinct source features and they do not have to follow the coherence and cohesion principles that define single documents."

In short, this approach brings comprehensiveness into the learning process because it requires that learners build interconnections and aim at a coherent understanding not only within a single text but across multiple texts (Britt & Rouet, 2012; Goldman, 2004).

Multiple text reading supports comprehension by integrating new information into pre-existing concepts and by grouping information into higher-order units, thus leading to a more profound and longer-lasting knowledge acquisition (Bråten et al., 2020; Britt & Rouet, 2012). Undoubtedly, this requires certain effort during the entire reading process and places high demands on the working memory (Bråten & Strømsø, 2011).

Much of the research on multiple-text reading focuses on conflicting-view documents (Castells et al., 2021), it explores how learners integrate conflicting information (e.g., Richter & Maier, 2017; Stang Lund et al., 2018) and assess the credibility of sources (e.g., Stadtler et al., 2013; Strømsø et al., 2013). However, it is important to note that within the approach of multiple text reading, there are instances where the emphasis has been placed on texts carrying semantically congruent or consistent information, as evidenced by examples from Braasch et al. (2016) and List et al. (2021). These studies deliberately utilized texts or incorporated parts of texts with consistent/congruent information. The connections that emerge when assessing the information presented in multiple complementary or congruent texts exhibit a continuum of characteristics. At one end, one has redundant or overlapping connections, where the same information is essentially restated or rephrased in a different manner across the texts. At the other end, one encounters distinct or componential connections, denoting a complementary relationship between the information found in one text and that which is contained in another, as each text contributes unique elements to the overall understanding (Castells et al., 2021).

To explain how the meaning is extracted from reading of multiple texts, several theoretical models were developed. The documents model (Britt et al., 1999; Perfetti et al., 1999; Rouet, 2006) explains how the mental representation from multiple texts might be created. Since this model originates in the integration–construction model (Kintsch, 1994, 1998), which describes how the meaning from a single text is extracted, the first three levels of this comprehension process are identical for both models. Namely, the first level, called the surface level, represents the level at which the exact wording and syntax of sentences are processed. At the second level, the text-base level, representation of the connections between sentences and parts of the text is created. The overall meaning of the text is represented at the third level, also known as the situation level. At this level, the fusion of text information and one's prior knowledge occurs. Then the documents model framework (Britt & Rouet, 2012) adds further level, the so-called integrated mental model, where integration of information from multiple texts occurs and additional mental representations about the concepts described in the texts are built. The variety of information present across the texts is connected and is fused into a global picture about the concepts and situations that are described in the texts (Britt & Rouet, 2012). Finally, the intertext model represents a mental representation of source specifics (authors, credibility of sources, etc.,) which also play an important role in multiple text synthesis (Bråten et al., 2020; Britt & Rouet, 2012).

Since extracting meaning from multiple texts is a very complex task with many underlying processes, many factors can impact these processes (McNamara et al., 2015). Primor and Katzir (2018) reviewed research on the multiple text reading approach and described and categorized the factors that play an important role in comprehension processes. The relationship between the texts to be read is considered to be a first influential factor. Using contradictory, complementary, or mixed information across the texts results in different outcomes. Another influential factor is the type of task to be performed after reading multiple texts. In this case, two categories were identified, either expressive tasks (essays writing, open ended questions) or receptive tasks (marking the correct response, verifying sentences). The way the results are assessed, either holistically or by analyzing smaller units, is also an important factor in this process (see also Barzilai et al., 2018). Finally, the information integration level at which participants are required to work plays an essential role. This factor differentiates whether participants are required to a) select important information, b) generate intertextual relationships, or c) make inferences.

Regarding the studies that compared multiple text reading to the reading of a single text (Nokes et al., 2007; Wiley & Voss, 1999), results show that reading multiple texts focusing on the same subject provides readers with a wider conceptual knowledge and a more integrated understanding. Nokes et al. (2007), who conducted their study with 246 high school students (aged 16–17) in history classrooms, found that multiple text reading enhanced students' understanding of historical content. Similarly, Wiley and Voss (1999), who worked with 64 undergraduate students, demonstrated that those who read multiple texts about historical content in preparation for writing an argumentative essay used more transformed sentences in their writing and resorted less frequently to copying sentences than learners who reread only one text. The present study also aims to examine the effectiveness of both multiple text reading and rereading, albeit under different conditions.

Metacognition

Metacognition refers to the monitoring, controlling, and assessment of acquired knowledge and one's progress toward own learning goals. Therefore, it allows for reflection of one's learning process and has an impact on decisions about the use and regulation of learning activities and techniques (Bjørk et al., 2013; Dignath et al., 2008; Dunlosky & Lipko, 2007; Schraw et al., 2006).

Learners' beliefs regarding attained knowledge can be measured in various ways. One of the most frequently used measures in research is the judgement of learning (JOL), which quantifies learner's subjective assessment of future performance. JOLs are made during learning and prior to testing and what is calculated is the accuracy of match between the anticipated and actual performance (Schraw, 2009). These metacognitive assessments can be accurate, in which case the learner makes appropriate adjustments during the learning, takes suitable decisions about strategies and techniques to be deployed, appropriately adjusts the learning schedule, etc. JOLs can also be inaccurate, which may result in counterproductive steps, such as premature termination of studying or choice of inappropriate learning techniques and strategies (Metcalfe, 2009; Nietfeld & Schraw, 2002; Rawson & Dunlosky, 2008).

Evidence suggests that these erroneous metacognitive assessments about one's knowledge take place because learners focus on the wrong cues when making their JOLs (Koriat, 1997; Rhodes & Castel, 2008; Thiede et al., 2010). Fluency, for instance, is a cue which during the learning process creates a sense of familiarity and evokes a subjective feeling that since the learning went smoothly, the material must be easy to remember (Koriat, 2008; Miele et al., 2011; Oppenheimer, 2008). As hinted above, rereading is one technique that leads to such a sense of familiarity (Wiley et al., 2005). However, it remains unclear how reading of multiple texts will impact the monitoring accuracy.

The present study

For an effective self-regulated learner, it is essential to be familiar with techniques and approaches that boost knowledge acquisition and long-term retention. These techniques and approaches should moreover promote accurate metacognitive monitoring of one's learning process (Bjørk et al., 2013). Two reading approaches will be in focus of this study—multiple text reading and rereading. Although both approaches involve reading and are easily accessible, they may appear more similar from a general perspective. However, learners often find multiple text reading more challenging (Rouet, 2006) due to the need to integrate information from various sources. This difference may influence learners' perceived effectiveness of each approach. According to the integration–construction model (Kintsch,), to comprehend a text one must build a third-level representation of it, namely a situation-level representation. What is important in this context is that, according to the dominant view, rereading of a text does not update this situation-level representation (Callender & McDaniel, 2009; Fritz et al., 2000; Kintsch, 1994). On the other hand, according to the documents model framework (Britt et al., 1999; Perfetti et al., 1999), multiple text reading enables to create an integrated representation of the topic which is supposed to be updated with each reading. This should lead to a more elaborate and durable knowledge acquisition.

In order to investigate our goals, the following adjustments to methodology commonly used by multiple text researchers will be made. First, since we aim to explore genuine effects of multiple text reading on comprehension and knowledge acquisition of naïve learners, we will refrain from using any learning techniques and strategies in connection with it (e.g., summarization, note-taking, highlighting). Second, as we aim to explore the effects of a multiple text reading approach compared to the most utilized technique—rereading (Dunlosky et al., 2013; Karpicke et al., 2009), we will employ reading materials that equally elaborate on given concepts. For that reason, we will refrain from using any contradictions in the texts, which is shown to have miscellaneous effects on learners (e.g., Barzilai & Eseth-Alkalai, 2015; Ferguson et al., 2013). Furthermore, as rereading shows its major benefits when distributed (Dunlosky et al., 2013; Rawson & Kintsch, 2005), we set to distribute learning sessions for multiple text reading as well. Third, since we are interested in exploring effects on comprehension, rather than on memorization, we will use open-ended questions as our assessment method (a type of expressive integration task, see Primor & Katzir, 2018). To sum up, we will attempt to research the effects of distributed reading of three different texts that contain congruent information without the use of any additional learning techniques as well as without any other features that multiple text reading approach contains (e.g., possibility to directly compare information across texts).

The aim of this study is to explore and assess the effectiveness of multiple text reading and rereading. In particular, we shall investigate participants' performance on a knowledge test focused on the depth of understanding and examine the accuracy of participants' metacognitive monitoring. This will be done both from a short-term (one day) and a long-term (one week) perspective. To the best of our knowledge, no study so far explored the effect of distributed reading of multiple texts without the use of additional study aids.

Hypotheses

Given that additional synthesis at an integrated level occurs during the reading of multiple texts, and that more elaborated representations of the concepts should be constructed (Britt et al., 1999; Perfetti et al., 1999), we hypothesize that participants in the multiple text reading group (hereafter MTR) will score higher on a knowledge test than participants in the rereading group (hereafter RR) in both short-term and long-term conditions (H1).

Participants in the experimental groups will read texts in a distributed manner and encounter key concepts repeatedly. Distributed learning has been shown to have beneficial effects on long-term knowledge retention (e.g., Bjørk et al., 2013; Dunlosky et al., 2013).

Based on this, we predict that participants in the multiple text reading (MTR) group and the rereading (RR) group will score higher than those in the single reading (1R) group on a knowledge test. This prediction applies to both short-term and long-term conditions (H2).

As forgetting is an inevitable process, both the experimental groups as well as the control group will naturally attain lower test scores in the long-term condition. However, we assume that in the case of MTR group, thanks to more elaborated representations of concepts that should be constructed (Britt et al., 1999; Perfetti et al., 1999), there will be no significant difference in test scores between the short-term and the long-term condition (H3a). On the other hand, in the case of rereading, even though research showed long-term benefits of distributed rereading compared to single reading (Rawson & Kintsch, 2005), it is presumed that there is no update of the representation of concepts during rereading (Callender & McDaniel, 2009; Fritz et al., 2000). Therefore, we assume that rereading does not provide learners with lasting knowledge and for that reason participants in the RR group will score significantly lower on the knowledge test in the long-term condition than in the short-term condition (H3b).

Considering monitoring judgements, rereading is a learning technique that provides learners with a misleading impression of confidence about one's acquired knowledge (Wiley et al., 2005), thus we assume that participants in the RR group will provide higher JOLs after each reading session than participants in the MTR group (H4).

There is little knowledge of the accuracy of monitoring judgments in multiple text reading approach. However, it is reasonable to assume that since participants in the MTR group will see each text only once, the subjective feeling of confidence will not arise. Therefore, we predict that participants in the MTR group will more accurately estimate their actual performance in the test (in both short- and long-term condition) than participants in the RR group in terms of delayed JOLs (measured before the final knowledge test) (H5).

Method

Open science practices

The project was pre-registered in the Open Science Framework (OSF). All materials and detailed description of the procedure (sample size calculation, exclusion criteria, etc.) can be found here: osf.io/hznsb

Participants and design

We conducted an a priori power analysis (multiple regression with interaction of two predictors) with the effect size set at medium (Cohen's $d=0.68$), alpha at 0.05, and the power of 0.90. Estimation of the effect size was based on two studies. Rawson and Kintsch (2005) compared single reading to rereading and arrived at a mean effect size Cohen's $d=0.78$, while a metanalysis of studies of cognitive learning strategies (Dignath et al., 2008) arrived at an effect size Cohen's $d=0.68$ for elaborative learning techniques on reading and writing performance.

In terms of our study, this would amount to a difference of five points in the means of scores achieved in the final test between the groups (with estimated $SD=7.4$ for each group). Based on this calculation in R, version 3.6.1 (R Core Team, 2019) using the pwr package (Champely, 2020), we needed at least 180 participants (30 participants in each of the 6 subgroups).

In total, 216 participants took part in the experiment. After applying exclusion criteria, 30 participants were excluded from further analyses—three participants who reported a reading impairment diagnosis, five who reported as their first language a language different from the language of the texts, four who did not take the final test, eight who correctly answered items on the prior knowledge test (see below), eight who failed to complete all three readings, and two who were not psychology students.

The final sample consisted of 186 participants (166 women, 20 men). Participants were homogenous in race and socio-economic status. All were students of psychology ($M_{age} = 20.76$, $SD_{age} = 2.03$) whose first language was the language of the texts. The study was approved by the Ethics Committee of University of Ss. Cyril and Methodius (reference number: FPV-21-2021).

Before the start of the experiment, participants signed an informed consent form and were informed that they were participating in a study focused on self-regulated learning mainly evaluating different reading approaches and their impact on learning. They were also informed that they could terminate their participation at any point without providing a reason. Participants were kept unaware of the specific hypotheses and detailed goals of the experiment to prevent bias. At the end of the experiment, participants had the chance to enter a lottery, in which 16 participants won a €30 bonus card for supermarket purchases.

The experiment was divided in two phases, a learning phase and a testing phase (see Fig. 1). For the experimental groups, learning took place on three consecutive days. Each

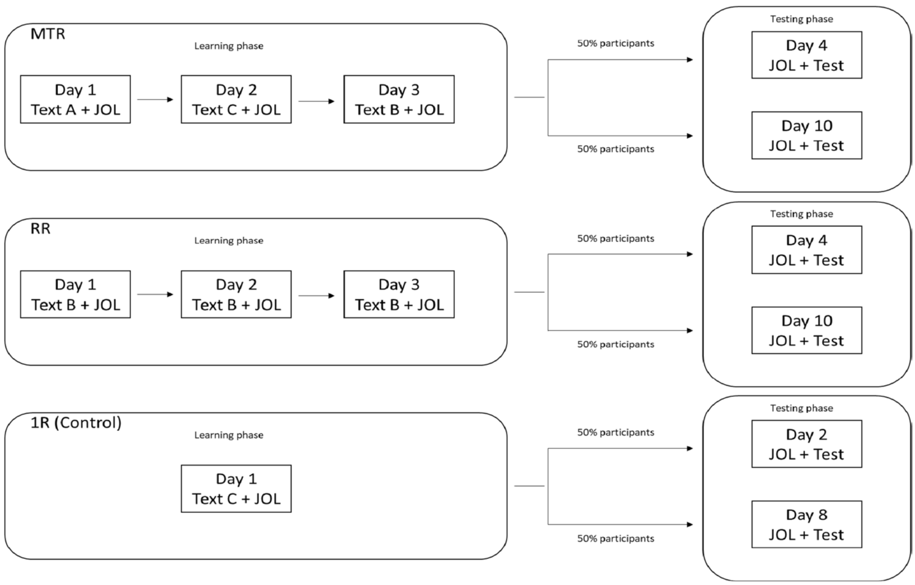

Fig. 1 A possible example of the experimental procedure for the three conditions. *Note:* Fig. 1 depicts only one of the possible scenarios of text order. For MTR group, participants could read texts in the following orders: ABC, ACB, BAC, BCA, CAB, or CBA. For RR group, participants could read AAA, BBB or CCC. For control group, participants could read text A, B or C. The order of the texts was semi-randomly assigned to ensure that all text combinations were used equally. MTR: multiple text reading; RR: rereading; 1R: single reading; JOL: judgement of learning

day, there was one reading session. For the control group, learning took just one day. After each reading, participants provided JOLs. At the beginning of the testing phase, participants provided delayed JOLs and then took the final test. Participants were randomly assigned to one of six subgroups defined by factorial combinations of two independent variables:

1. Reading approach:
   a. Multiple text reading (MTR)—each learning session involved reading a different text on the same subject;
   b. Rereading (RR)—each learning session involved reading the same text;
   c. Single reading (1R)—having only one reading session (control condition).
2. Time of the test:
   a. One-day delayed (DD)—taking a test one day after the last reading;
   b. One-week delayed (WD)—taking a test seven days after the last reading.

Texts were distributed to participants in a semirandom fashion to ensure that all three texts were used an equal number of times. Meaning that participants could receive texts in various permutations. In the multiple text reading group, it could be ABC, ACB, BAC, BCA, CAB, or CBA; in the rereading group it could be AAA, BBB, or CCC; and in the control group it could be A, B, or C. Participants were not exposed to knowledge test repeatedly, but we used different groups of participants, resulting in 3×2 factorial design. This decision was driven by the concern that the testing effect could overshadow the differences between the experimental groups.

Materials

Assessment of prior knowledge

As prior knowledge about the topic strongly influences the learning outcomes (Alexander & Judy, 1988), to eliminate this confounding factor, we decided to exclude such participants from our study. Learner's prior knowledge was assessed by a short test that was administered at the beginning of the experiment together with a questionnaire collecting demographic data. It consisted of three open-ended questions that scanned the general knowledge of participants about the topic to be learned (e.g., "What do you know about Dual-process theories of cognition?"). If a correct answer was provided to at least one of these questions, such participant was excluded from the study.

Texts

The experiment used three texts; all were scientific texts in the first language of the participants. Any typographical cueing in the texts was avoided. Each of the texts explained the principles of the dual process theory, heuristics in general, and anchoring heuristic and confirmation bias (theories and constructs proposed by Evans, 2003; Kahneman, 2011; Stanovich, 2004; Tversky & Kahneman, 1974; Wason, 1960). Structurally, each of these three texts was divided in six subsections. Each text was written by a different author;

information was thus presented in a different way in terms of style, wording, and organization of the text. Importantly, all texts had congruent content with respect to the conceptual information they carried, and all explained the key concepts to the same degree (see Appendix A for samples of the texts). All these materials were primarily intended for study purposes and were available in textbooks or on educational websites. Each text was 1,002–1,008 words long. The Flesch–Kincaid grade-level scores for the texts were 15.5–16.8, which indicates that the texts were intended for college students. Additionally, the texts had been assessed as equally demanding and content-equal by five independent experts. Two linguists assessed the structure, clarity of phrasing, and overall readability of the texts with a conclusion that texts are comparable in these aspects. Three cognitive psychologists concluded that key concepts are adequately explained in each text and that readers should be able to answer all test items by reading whichever of three texts. Also, after data were collected, we conducted one-way ANOVAs with the text sample (A, B, C) as the predictor variable and score on the final test as the outcome variable. Our aim was to test whether any of the texts allowed participants to perform better. We found no main effect of the text sample (for rereading: $F(2, 59) = 0.36$, $p = 0.70$; for controls: $F(2, 59) = 0.70$, $p = 0.50$).

Judgements of learning (JOLs)

We used JOLs to measure metacognitive monitoring during learning and before performance in the knowledge test. In the learning phase, participants were asked: "*If you were to take the test with 15 open-ended questions that would map your knowledge of the subject you studied during these sessions, how many points would you score? The maximum that can be obtained is 30 points. For each question, you can receive two points for a fully correct response, one point for a partly correct response, and no points for an incorrect response.*"

In the testing phase, delayed JOLs were measured by asking students: "*How many points do you think you are going to score in the upcoming test consisting of 15 open-ended questions which map your knowledge of the subject you studied during the previous sessions?*" and, as before, provided explanation of scoring.

Since we were interested if participants in experimental groups are underconfident or overconfident about their performance, to assess their predictive accuracy, we have computed a bias index. The bias index is computed as follows: From the confidence rating (JOL) the actual score attained on the test is subtracted and then it is divided by the maximum number of points that is possible to gain on the test. Final number ranges from -1, indicating underconfidence of a participant, to 1, indicating overconfidence of a participant, with 0 indicating a perfect accuracy of a prediction (for details, see Schraw, 2009).

The knowledge test

The knowledge test consisted of fifteen open-ended questions which participants should have been able to answer based on the reading of either of the three texts. Answers had the form of explanations of key concepts used in the texts (see Appendix B for examples of questions).

Scoring: Participants could obtain a maximum of two points for each answer. The maximum score on the test was 30 points. The first and second authors graded the responses, both of whom have backgrounds in cognitive and educational psychology. Both authors were blinded to the condition. The research assistant randomized the documents and removed the column indicating the experimental condition. Prior to assessing participant responses, authors carefully crafted a scoring rubric detailing the criteria necessary for awarding two points, one point, or zero points for each answer. This rubric was designed to ensure consistency and objectivity in grading across all experimental conditions. Importantly, the scoring rubric was structured to allow participants to provide comprehensive answers by drawing from any of the provided texts. This approach aimed to mitigate potential biases introduced by variations in information across the texts. Then, authors independently scored all test sheets. Finally, they compared their scoring and examined and resolved any inconsistencies in the scores through discussion.

To summarize the principles of the scoring rubric with an example: Two points were awarded for a fully correct response where understanding was evident and the answer was sufficiently comprehensive. For instance, for the question item: *"Explain in a few sentences the principle on which the 'analytical' System 2 works,"* a two-point answer would be: *"System 2 operates on the principle of logic and thinking. It is much more accurate and objective compared to System 1, but it requires a significant amount of energy to function. It takes more than one answer into account and tries to evaluate situations from different perspectives."* Another example of a two-point answer could be: *"This system is characterized by the fact that we have to consciously and 'manually' draw out and process or link our experiences to new information. System 2 is not automatic, meaning we need to be mentally engaged in the situation. This is called slow thinking."*

One point was given for a partially correct response that showed some understanding but was either incomplete or excessively general. For example, the same question might be answered with: *"I remember it was called 'slow thinking', involving thinking and dissecting a topic, finding a way out,"* or *"Unlike the first system, this system is will-controlled, so opinions and thoughts are more detailed."*

Zero points were given for an incorrect response or no response at all. Examples of zero-point answers include: *"It is a system based on rules,"* or *"It is slow."*

Procedure

The experiment was completed on-line under experimenter's supervision. Participants were logged-in via one of the video communication platforms. The participants and the experimenter maintained direct online contact throughout the experimental session.

All participants were instructed that they would go through a learning phase and subsequently be tested on the knowledge thus acquired with open-ended questions, which would require understanding and ability to explain concepts. Participants in the experimental groups were told that there would be three learning sessions over three consecutive days. However, they remained naïve to which experimental group they would be assigned as well as when the final test would take place. Also, no information about the nature of the texts was provided. Participants in the control group were informed that only one learning session would take place. After completion of the last reading, half of participants from each condition were instructed to log in to a scheduled online meeting and take a knowledge test on the day after the last reading; the other half were instructed to log in for a test seven days after the last reading session. Participants were asked to avoid any material related to the subject of the texts and to mentally rehearse what they had learned. They were also instructed to refrain from the use of any other learning technique during the reading (such as highlighting, writing of summaries, etc.).

Before starting the reading, participants completed a questionnaire on demographic information and prior knowledge. Then they were instructed to open a link in the chat and read a text in a self-paced manner as if they were reading for an exam. They were encouraged to take their time and ensure they understood the material thoroughly. Average time of reading was 15 min and 47 s. After reading, participants provided their estimates (JOLs) of how many points they would score if the final test were taken right at that point.

For the testing session, participants logged in to an online meeting at a pre-scheduled time. They received a link to the final test, which was conducted online. The final test was limited to 45 min and consisted of fifteen open-ended questions. Before starting the test, participants provided their estimates of performance (delayed JOL) based on their confidence in the material they had studied. Participants were required to proceed through the test questions sequentially and could not revisit previously answered questions.

Data analyses

All analyses were conducted in R version 3.6.2 (R Core Team, 2019) in an integrated development environment R Studio (RStudio Team, 2020). We used the following R packages: car version 3.1–0 (Fox & Weisberg, 2019), dplyr version 1.0.2 (Wickham et al., 2020) and emmeans version 1.7.1–1 (Lenth, 2021), and mirt version 1.33.2. (Chalmers, 2012).

Results

Descriptive statistics (means and SDs) for measured variables are displayed in Table 1. Cronbach's α for the one-day delayed knowledge test was $\alpha_{MTR} = 0.91$, $\alpha_{RR} = 0.89$, $\alpha_{1R} = 0.82$ and Cronbach's alpha for the one-week delayed knowledge test was $\alpha_{MTR} = 0.91$, $\alpha_{RR} = 0.88$, $\alpha_{1R} = 0.78$. In assessing the validity of our test, we used the simple Graded Response Model. The results of this analysis revealed that the test items effectively differentiated between participants with varying levels of latent ability. We identified three specific question items that were notably more challenging than others, making it more difficult for participants to achieve the maximum score on these items (Appendix C). The test was found to be most informative of respondents with latent scores ranging from -1.5 to 1.5, effectively targeting participants with average abilities. However, the overall model fit was not entirely satisfactory ($M^2 = 309.386$, df = 75,

Table 1 Means and SDs of measured variables

Approach	Time of the test	Score	SD_{score}	dJOL	SD_{dJOL}	Bias Index	$SD_{Bias\ Index}$
MTR	DD	15.5	8.88	11.7	5.42	−0.13	0.27
MTR	WD	12.9	8.89	8.8	5.92	−0.14	0.23
RR	DD	14.1	7.91	16.6	6.08	0.08	0.22
RR	WD	10.1	7.64	10.5	4.94	0.01	0.19
1R	DD	12.0	6.82	11.8	4.33	−0.01	0.25
1R	WD	4.3	4.43	9.2	5.77	0.16	0.22

MTR multiple text reading, *RR* rereading, *1R* single reading, *DD* one-day delayed, *WD* one week delayed, *Score* number of points obtained in the knowledge test, *dJOL* judgement of learning collected right before the knowledge test

$p < 0.001$, RMSEA = 0.130, SRMSR = 0.110, TLI = 0.856, CFI = 0.880). The SRMSR index suggested the presence of high residuals between items, prompting us to investigate local item dependencies. We applied Q3 statistics to test for local dependency between pairs of items. It became apparent that several pairs of items displayed significant item dependency, which could be attributed to the test's structure that includes clusters of items related to common stimuli, known as testlets. We identified 4 testlets. Consequently, we applied the testlet version of the Graded Response Model, which resulted in a satisfactory model fit ($M^2 = 165.785$, df = 89, $p < 0.001$, RMSEA = 0.068, SRMSR = 0.071, TLI = 0.969, CFI = 0.974).

Correlations between measured variables can be found in Table 2.

Before fitting the main models, we have performed a one-way ANOVA with a JOL on Day 1 as an outcome variable to test whether there was any significant difference between groups on Day 1. The reading approach was a predictor. This ANOVA did not reveal any significant difference between the groups, $F(2, 183) = 0.51$, $p = 0.60$.

Table 2 Correlations between measured variables

	J1	J2	J3	dJOL	Score
A. Multiple text reading group					
One-day delayed test (bottom triangle, n = 32) and One-week delayed test (upper triangle, n = 30)					
J1	–	.73***	.67***	.71***	.48**
J2	.65***	–	.70***	.65***	.46**
J3	.71***	.75***	–	.76***	.65***
dJOL	.64***	.70***	.67***	–	.63***
Score	.44**	.23	.43**	.43**	–
B. Rereading group					
One-day delayed test (bottom triangle, n = 32) and One-week delayed test (upper triangle, n = 30)					
J1	–	.84***	.69***	.71***	.50**
J2	.89***	–	.86***	.57***	.40*
J3	.68***	.88***	–	.51**	.47**
dJOL	.61***	.74***	.79***	–	.66***
Score	.68***	.76***	.55**	.60***	–

	J1	dJOL	Score
C. Single reading group			
One-day delayed test (bottom triangle, n = 32) and One-week delayed test (upper triangle, n = 30)			
J1	–	.59***	.17
dJOL	.57***	–	.17
Score	.37*	.14	–

MTR multiple text reading, *RR* rereading, *1R* single reading, *J1* judgement of learning collected after the first reading, *J2* judgement of learning collected after the second reading, *J3* judgement of learning collected after the third reading, *dJOL* judgement of learning collected right before the knowledge test; Score: number of points obtained in the knowledge test

*$p < .05$. **$p < .01$. ***$p < .001$

Analyses of performance

We have conducted a two-way ANOVA with the test score as the outcome variable and the reading approach in interaction with the time of the test as predictors. To check for the assumptions of normally distributed and homogeneous residuals, we have visually inspected the QQ plot and a scatterplot of residuals plotted against the fitted values (Field, 2005); this did not indicate any clear deviations from these assumptions. Next, we assessed the model's stability using DFBeta-values and Cook's distance (Field, 2005); this showed the model to be stable. We have therefore concluded that the statistical model provides an adequate representation. There was a significant main effect of the reading approach that was used during the learning phase on the performance on the knowledge test, $F(2, 180) = 9.81$, $p < 0.001$, $\eta p^2. = 0.10$, and also there was a significant main effect of the time of the test on the performance on the knowledge test, $F(1, 180) = 18.03$, $p < 0.001$, $\eta p^2 = 0.09$. There was not a significant interaction between the reading approach and a time of the test on the performance on the knowledge test, $F(2, 180) = 1.90$, $p = 0.15$, $\eta p^2 = 0.02$.

To answer the set of hypotheses related to the performance on the knowledge test (H1 – H3), we conducted a post-hoc analysis based on the Tukey HSD adjustment for pairwise comparison. For Hypothesis 1, which predicted that MTR group will perform better than RR group on the test taken one day after last reading as well as on the test taken one week after last reading, we detected nonsignificant differences between the groups on the one-day delayed test as well as on the one-week delayed test. These outcomes show that multiple text reading and rereading lead to similar results, with multiple text reading delivering slightly better results especially in the one-week delayed test (see Table 1). For Hypothesis 2, which predicted that MTR and RR groups will score higher both one day after last

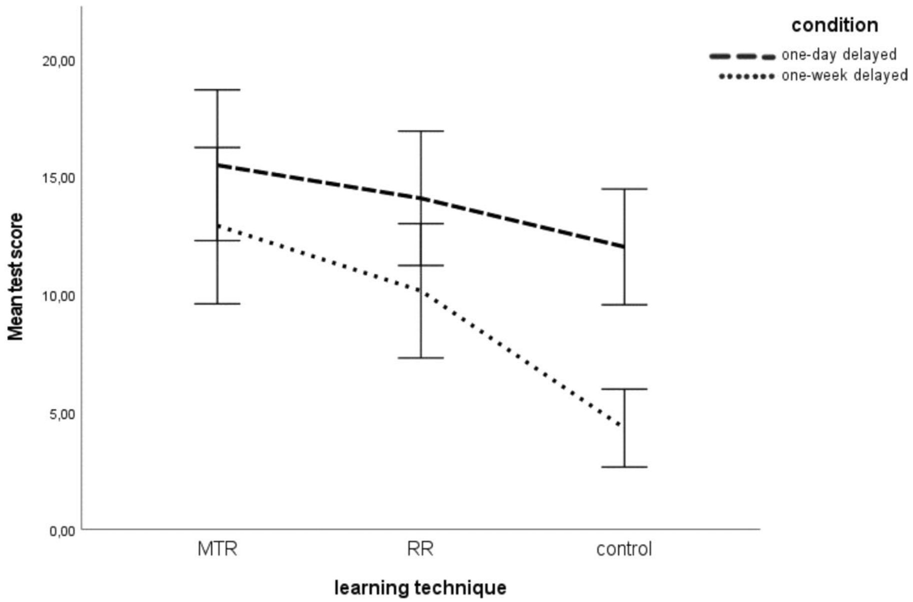

Fig. 2 Mean scores achieved in the knowledge test. *Note*: Attained score on the knowledge test as a function of employed reading approach (MTR: multiple text reading; RR: rereading; 1R: single reading) and time of the test (one-day delay or one-week delay between study and the test). Maximum score was 30 points

reading and one week after last reading than the 1R control group, we found nonsignificant differences in the means between the reading approaches and the control group on the one-day delayed test, while on the one-week delayed test, differences between the experimental groups and the control group were significant. For differences between MTR group and control group, it was $p<0.001$, 95% C.I. = [− 14.25, − 2.95] and for differences between RR group and control group, it was $p=0.039$, 95% C.I. = [− 11.48, − 0.18]. This indicates that multiple text reading and rereading yield comparable results as a single reading on the one-day delayed test but better results on the one-week delayed test (see Table 1). In terms of Hypothesis 3, a comparison of differences within both reading approaches in the scores achieved in the one-day delayed test versus one-week delayed test showed no significant differences. As predicted, though, for multiple text reading, the difference between the two testing points were smaller than for the rereading approach (for details see Table 1). This indicates that multiple text reading delivers a more stable and lasting knowledge than rereading does. The results are visually reported in Fig. 2.

Analyses of JOLs

To test Hypothesis 4, which predicted a higher overconfidence in the RR group than in the MTR group, we conducted a two-way mixed ANOVA with JOLs collected after each reading (three times) as a within-subject factor and the reading approach (without the controls) as a between-subject factor. The results showed that there was a significant main effect of the timepoint of measuring JOLs: $F(2, 244)=34.03$, $p<0.001$, $\eta^2=0.22$, and also a significant main effect of the reading approach $F(1, 122)=4.64$, $p=0.03$, $\eta p^2=0.04$. Finally, there was a significant interaction between the time-points of measurement of JOLs and the reading approach that was utilized: $F(2, 244)=9.54$, $p<0.001$, $\eta p^2=0.07$. For the MTR group, post-hoc test based on Tukey HSD for pairwise comparison showed a significant difference between JOLs after the first and the second reading ($M_{\mathrm{JOL1}}=14.73$, $SD_{\mathrm{JOL1}}=6.08$; $M_{\mathrm{JOL2}}=16.26$, $SD_{\mathrm{JOL2}}=5.14$, $p=0.034$, 95% C.I. = [− 0.04, − 0.02]) and no difference in JOLs after the second and the third reading ($M_{\mathrm{JOL2}}=16.26$, $SD_{\mathrm{JOL2}}=5.14$; $M_{\mathrm{JOL3}}=16.16$, $SD_{\mathrm{JOL3}}=6.66$). For the RR group, there was a significant difference in the JOLs between the first and the second reading ($M_{\mathrm{JOL1}}=15.45$, $SD_{\mathrm{JOL1}}=5.11$; $M_{\mathrm{JOL2}}=17.66$, $SD_{\mathrm{JOL2}}=5.02$, $p<0.001$, 95% C.I. = [− 3.72, − 0.70]) as well as between the second and third reading ($M_{\mathrm{JOL2}}=17.66$, $SD_{\mathrm{JOL2}}=5.02$; $M_{\mathrm{JOL3}}=19.89$, $SD_{\mathrm{JOL3}}=5.09$, $p<0.001$, 95% C.I. = [− 3.74, − 0.72]). In short, rereading led to anticipation of higher scores after each reading, while multiple text reading resulted in higher JOLs only after the second reading – the third reading did not make leaners any more certain about their performance. Figure 3 graphically reports the results.

For Hypothesis 5, which predicted that delayed JOLs of the MTR group would more accurately estimate the performance in the knowledge text than delayed JOLs of the RR group, we have first computed a bias index from JOLs acquired before the final test and the final score on the test (Schraw, 2009). Next, a two-way ANOVA was conducted with the bias index as the outcome variable and reading approach and time of the test in interaction as predictor variables. This revealed a significant main effect of the reading approach ($F(2, 180)=14.54$, $p<0.001$, $\eta p^2=0.14$), a non-significant main effect of time of the test ($F(1, 180)=0.77$, $p=0.38$, $\eta p^2=0.00$), and a significant effect of interaction ($F(2,180)=4.44$, $p=0.013$, $\eta p^2=0.05$). Subsequent post-hoc test (Tukey HSD correction for pairwise comparison) showed a number of differences between the confidence of learners: JOLs

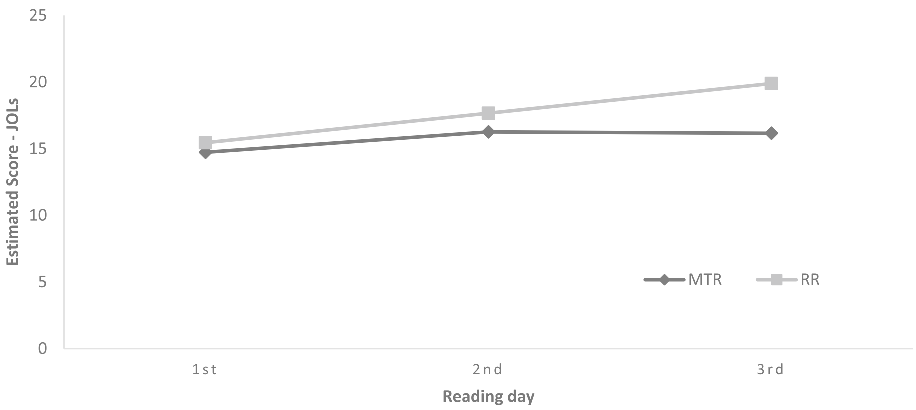

Fig. 3 A line graph showing changes in the JOLs after each reading. *Note*: Estimated score (JOLs) as a function of the reading day (first, second, third) and the employed reading approach (MTR: multiple text reading; RR: rereading)

collected one day (DD) after the last learning session showed that learners in the multiple text reading group (MTR) were underconfident about the knowledge they acquired while those in the rereading group (RR) were overconfident, and the difference was significant ($p<0.01$, 95% C.I. = [0.04, 0.38]). In delayed JOLs collected on a one-week (WD) delayed test, there was a nonsignificant trend in differences between these two groups, because learners in the rereading group became more accurate ($p=0.13$, 95% C.I. = [−0.02, 0.32]). This indicates that multiple text reading in general prompts students to being underconfident about the knowledge they acquired, while rereading does the opposite.

We have also found a significant difference in the bias index on the one-week (WD) delayed test between the multiple text reading group (MTR) and the single reading group (1R): learners in the 1R group did not expect they would have forgotten as much as they did and were overconfident in their estimates ($p<0.001$, 95% C.I. = [0.13, 0.47]). The means of the groups are presented in Table 1.

Discussion

This study investigated whether distributed reading of multiple texts without the support of any other learning techniques is effective for studying and acquiring lasting knowledge. Multiple text reading was compared to both rereading and single reading. Acquired knowledge was measured by 15 open-ended questions either one day or one week after the last reading session. On top of that, we measured metacognitive monitoring of reading comprehension using JOLs collected after each reading session and before the test.

When the knowledge test was administered one day after the last reading session, a comparison of results of multiple text reading, rereading, and single reading showed only small and nonsignificant differences between the groups. When, however, the test was administered one week after the last reading session, the multiple text group performed the best among the groups, while the rereading group performed significantly better than the single reading group. Multiple text reading may support better knowledge retention, as suggested by the small changes in scores between the one-day and one-week delayed tests.

Regarding the accuracy of metacognitive judgements which were gathered after each reading, participants in the multiple text reading group showed increased confidence only after the second reading, while participants in the rereading group were significantly more confident about their future performance after both the second and the third reading. Aside from this, we have also computed a bias index from the performance on the test and delayed JOL (gathered right before the test). Its results show that participants in the multiple text reading group were underconfident about their performance in the upcoming test in both the one-day delayed and one-week delayed condition. The rereading group, on the other hand, showed more calibrated estimates of their performance, slight overconfidence regarding the one-day delayed test and accurate assessment of the one-week delayed test.

Performance in the knowledge test

Contrary to our initial predictions our data did not show any significant differences between multiple text reading, rereading, and single reading in the one-day delayed test. Although the single reading group achieved the lowest scores, the effect of either multiple text reading or rereading was not as high as one would expect in view of the time and effort invested in learning. In the one-week delayed test, though, the differences between the groups were pronounced. Both the multiple text reading and rereading group performed significantly better than the single reading group. A comparison between the performance of the multiple text reading and rereading group in the one-week delayed test showed that participants in the multiple text reading group achieved higher scores, but this difference was not formally significant. It thus did not confirm our hypothesis, quite possibly because we sought a medium-sized effect (based on a metanalysis by Dignath et al., (2008), which showed that the effects of elaborative and organizational learning techniques are of a medium size). One could speculate that if the trend we observed in our data continued and we were trying to find a small-sized effect, we would have found it.

These overall results are in line with the construction–integration model of representation of textual sources which introduces three levels of representation: the surface level, text-based level, and the situation level (Kintsch,). According to this model, text comprehension requires a situation-level representation, which allows for a longer retention of information than representations made at lower levels, such as for instance memorization of word lists (see also Fisher & Radvansky, 2018; Kintsch et al., 1990; Radvansky et al., 2001). Since our test did not require the participants to memorize information but to understand and explain certain concepts, they had to build a situation-level representation (Kintsch, 1994). Given that such representation can be constructed even after a single reading, even the control group participants were able to explain the key concepts embedded in the text and perform rather well in the one-day delayed test; in this test, neither reading of multiple texts nor rereading led to significantly better results.

The situation changes when one needs to retain the situation-level representation for a longer period. In such an instance, repetition is needed. In the one-week delayed test, the single reading group thus performed significantly worse than either the multiple text or rereading group. This indicates that after seven days, the situation-level representation acquired through a single reading is not sufficiently strong and stable. This is supported by the findings of Fisher and Radvansky (2018), who found that after a single reading, situation-level memory changes around seven days after its formation and it is around this timepoint that it starts to quickly fade and deteriorate.

In both the multiple text reading and rereading groups, where information was rehearsed multiple times, the forgetting process appeared to be slower compared to single reading. However, the two reading approaches resulted in slightly different patterns of performance over time. Although the differences were not statistically significant, the multiple text reading group showed better results in the one-week delayed test compared to the rereading group. Additionally, the decline in scores between the one-day and one-week delayed tests was smaller in the multiple text reading group. This suggests that multiple text reading may contribute to more stable knowledge retention over time, whereas rereading could lead to a more rapid decline in the strength of the acquired representation.

The difference in the results of multiple text reading and rereading can be explained in terms of two models: the abovementioned construction–integration model (Kintsch, 1994, 1998) and the documents model (Britt et al., 1999; Perfetti et al., 1999). This documents model builds upon the construction–integration model and adds another level on which the integrated mental model is built. This helps create a more elaborate representation and allows for updating the situation-level representation, which is the level that enables learners to extract meaning from the text (Kintsch, 1994). In rereading, the dominant view is that the situation-level representation does not update after each reading (see e.g., Callender & McDaniel, 2009; Willey et al., 2005; Stine-Morrow et al., 2004; but cf. Millis et al., 1998). The documents model, however, proposes that in multiple text reading, the situation-level representation is updated via intertextual synthesis, which leads to more lasting knowledge (e.g. Bråten et al., 2020; Britt et al., 1999). This explanation can be supplemented by considering the distributed practice effect, which is often stronger in delayed tests than in the immediate ones (Dunlosky et al., 2013). Therefore, even if rereading does not lead to updates of the situation-level representation, distributed practice itself allows for the retention of certain amount of knowledge. In contrast, the control group, which read the text only once, did not benefit from distributed practice. As a result, while there were nonsignificant differences between the reading approaches and the control group on the one-day delayed test, both multiple text reading and rereading led to significantly better outcomes than the control group on the one-week delayed test. This suggests that the additional exposure to the material in the multiple text reading and rereading groups contributed to better long-term retention compared to a single reading.

Relating to the research conducted by Wiley and Voss (1999), it is evident that a direct comparison is complicated due to differences in methodologies. However, it is worth noting that despite these methodological differences, there are similarities in the results which hint towards the benefits stemming from using multiple texts. Wiley and Voss (1999) conducted two experiments aimed at enhancing undergraduate students' understanding of historical subject matter. In their study, students acted as historians, constructing their own models of historical events. Participants were provided with information from multiple sources on a website, rather than a single textbook, and were instructed to write argumentative essays. Researchers were mainly interested in the quality and characteristics of the responses they received. Moreover, participants had an unrestricted access to reference materials during the composition process and were given varying set of instructions under which they were supposed to craft their essays. They also manipulated with different instructions on how to structure their essays. On the other hand, our research was designed to reflect the typical learning experiences of contemporary learners. We intentionally crafted a more straightforward experimental setup, with a focus on information comprehension and retention. We used a knowledge test with open-ended questions to assess whether participants could correctly explain important concepts. While some responses were more detailed than others, we primarily focused on whether the answers contained the necessary information, regardless of their level of detail, i.e., the quality and elaboration of answers were not analyzed.

In summary, the findings suggest that reading from multiple texts not only allows for more elaborated answers with transformed and novel sentences on essays as found by Wiley and Voss (1999), but also yields discernible benefits in terms of long-term retention of conceptual knowledge.

Metacognitive monitoring (JOLs)

When students prepare for a test, they monitor their text comprehension. Information from monitoring has an impact on the regulation of further learning. Students in our study monitored their comprehension after each reading session and before the final test by providing an estimate of scores they would obtain in the final test.

After the first reading, we found no significant differences in estimates of scores between the three groups. Participants in both experimental groups increased their estimates after the second reading, with the rereading group showing a slightly higher increase than the multiple text reading group. The key difference took place after the third reading session. While participants in the rereading group significantly increased their estimates of future test performance, participants in the multiple text reading group did not increase their estimates at all. This shows that by seeing the same text repeatedly, participants in the rereading group gained more confidence after each reading than participants in the multiple text reading group did. These results are in line with the findings of Koriat (1997), who pointed out that learners in their estimates rely too much on experience-based cues, which are unfortunately misleading. During rereading, learners experience perceptual fluency, which is one of these experience-based cues. It brings about a sense of familiarity, which is erroneously interpreted as improved understanding of the text. Higher JOLs reflect this impression (see also Willey et al., 2005; Rawson et al., 2000). In the multiple text reading group, we did not find such increase in confidence because learners encountered each text only once (although it concerned the same concepts), there was no space for perceptual fluency to emerge.

Regarding the delayed JOLs measured before the test and the subsequently computed bias index, our results were contrary to our hypothesis and somewhat unexpected. Participants in the multiple text reading group were significantly underconfident about their performance in both the one-day and one-week delayed test, while participants in the rereading group were slightly overconfident about their performance in the one-day delayed test but their assessment became more accurate regarding the one-week delayed test. The lack of confidence observed in the multiple text reading group was likely due to the fact that, having encountered each text only once, they did not experience a sense of fluency. Because learners in their estimates rely heavily on experience-based cues and these were missing, they had no reason for overconfidence. From this perspective, multiple text reading could be viewed as a reading approach that introduces “desirable difficulties” into the learning process (Bjørk, 1994; Bjørk & Bjørk, 2011). Our results show that learners do not experience the fluency and ease of learning (reflected by low JOLs) but the actual performance is higher than they anticipate.

Instructional implications

Our findings have certain implications for the educational context. It seems that for average students who are not trying to acquire long-lasting knowledge, one reading on the day before the exam suffices for achieving results not much worse than those achieved by rereading or multiple text reading, especially if what is tested is comprehension (situation-level representation). But for long-term retention, it is more beneficial to read different sources on the subject over a period of time. This seems to result in a more integrated and elaborate representation, which in turn leads to longer lasting knowledge.

These findings also have important implications for self-regulated learning. Firstly, in cases when the accuracy of metacognitive monitoring of learning is in question, learners who study from multiple texts do not become overconfident. One could even assume that if learners who used the multiple text reading approach had a chance to review the material needed for an exam, they would do so (see also Dunlosky & Ariel, 2011; Dunlosky & Rawson, 2012; Metcalfe & Finn, 2008; Nelson et al., 1994), which could again benefit their test performance. Furthermore, as Lee and List (2023) found in their study, one could assume that students who utilize multiple texts will increase their help-seeking behaviors in order to comprehend the given topic (for a detailed explanation about the role of self-regulated learning in multiple text reading, see also Rouet & Britt, 2011).

The results of assessment of test scores in the light of metacognitive monitoring (collected JOLs) show that multiple text reading brings into the learning process desirable difficulties. Learners who used this approach were underconfident about their performance, which indicates that they felt they had not learned enough although, in fact, they outperformed the other groups in the delayed test. This has important implications for real-life learning. Students are often reluctant to use approaches that require more effort and time (Baars et al., 2022), they opt for easier and less time-consuming methods, first of all rereading. This study shows that for a person learning about a new subject, reading different texts about the same topic over a period of time leads to a longer retention of knowledge and is thus more beneficial than rereading.

Limitations

From the perspective of percentual success of participants, the knowledge test was difficult. The best-performing group, the multiple text reading one, achieved on the one-day delayed test only around 50% of points. The test effectively discriminated between participants with average performance, and even the best-performing students had the difficulties to respond to three items. We speculate that if the most difficult items were easier, it may have more sensitively highlighted the differences between the groups and testing points. But the percentual success of participants could have also been different had they been tested immediately after the last reading.

Our study sample predominantly consisted of females, comprising 89% of the participants, hence it remains a possibility that gender could have had potential influence on the results. Additionally, our sample was intentionally drawn exclusively from the pool of psychology students, aligning with our aim of engaging readers with a vested interest in the topic at hand. However, this deliberate choice may potentially limit the generalizability of our findings to a broader population.

The consideration of using repeated measures for participants' performance on the test is a valid alternative that we contemplated during the experiment's design phase. Nevertheless, we were concerned about the potential influence of the testing effect (Dunlosky et al., 2013; Rowland, 2014) on the observed differences between our experimental groups, basically that it potentially can mask the true effect of the reading approaches we sought to investigate. Given this concern, we opted for the approach of using different groups of participants across the study, thus implementing a 3×2 between subject design. While this approach may have its limitations in terms of generalizability and the potential influence of participant characteristics like intelligence, motivation, interest in the topic, etc., we considered these variables potentially less confounding than the testing effect.

In our experimental design, participants were randomly assigned to groups. But if we were exploring the real study preferences of participants and their reading capabilities, we might have found that multiple text reading has a different impact on different groups of learners. In particular, students who deliberately opt for the multiple text reading approach might differ in various respects from students who choose rereading and highlighting. They could, for instance, differ in the number of techniques they deploy, in their motivation to study, or even in their reading skills.

As we have described throughout the study, our approach involved an unconventional method of multiple text reading, deliberately avoiding any inherent contradictions within our texts. Moreover, our study did not involve participants in assessing the sources of the texts. Consequently, the scope for comparing our results with those of other studies remains somewhat constrained. Nevertheless, when we consider the text-belief consistency effect (Richter & Meier, 2017), it becomes apparent that learners often tend to disregard conflicting information and instead focus their attention on information that aligns with their preexisting beliefs. This suggests that our results may not be significantly divergent from the outcomes observed in earlier research.

Future directions

Given that our findings regarding metacognitive monitoring show that participants in the multiple text reading group underestimated their performance and participants in the rereading group were more confident about their performance after each reading, future research should try to investigate how students would decide regarding continuation of learning in the light of their perceived progress. For instance, whether participants in the rereading group would prefer to read less and not study for the test anymore and whether participants in the multiple text reading group would review the topic in question or opt for another method. Moreover, it would be interesting to measure perceived difficulty during the reading and see how the perceived difficulty influences the JOLs.

Appendix A

Examples of the text excerpts (These texts have been translated into English from the original language of the study).

Sample of the text A

In the era of research on rational-logical models of decision making, theories seeking to explain the decision-making process in humans have emerged. One important branch of research were the so-called dual-process theories of thinking. According to Kahneman (2012), when we think about ourselves, we tend to identify with System 2, our conscious, rational self. This System has formed certain opinions, it also makes various choices and decides what we should think and what we should do. Although System 2 believes that it is the one that decides how we behave and that it has the final say, it is mistaken, as System 1 plays a significant role in our decision-making. It is automatic; we can think of it as impressions, feelings and sympathies. It represents our innate skills within the psyche. It makes associations between ideas and can pull knowledge and information from memory without conscious effort. System 1 is also called "fast thinking". Kahneman goes on to describe that System 1 also has its limitations as it operates outside of our conscious control and runs continuously. By being on all the time, it tends to make mistakes, as it selects only some of the many stimuli acting on us to which it pays attention and distorts or generalizes the rest. Many times, it is difficult to leave the platform of quick thinking, as it is not regulated by the will and gives automatic answers. On the other hand, an important characteristic of System 2 is precisely that it is will-controlled. However, this requires sufficient mental commitment to be able to express opinions that are well thought out and explicitly logically grounded. System 2 is also referred to as "slow thinking."

Sample of the text B

Evans (2003) focuses on a detailed description of both systems of thinking, provides evidence in favor of the existence of dual processes and also describes individual differences in people's ability to reason. He characterizes System 1 as a universal form of cognition, common to both humans and animals. He does not see it as a unified system, but as a kind of set of subsystems that function to some extent autonomously. This system includes instinctive behaviour, which is genetically programmed. The processes of System 1 are fast, parallel and automatic. Only their final product is ever sent to consciousness. In contrast, system 2 is evolutionarily younger and is exclusive to humans. It allows hypothetical thinking and reasoning at an abstract level. However, it is limited by working memory capacity. According to the author, it is also correlated with a measure of general intelligence. Thinking in this system is slow and sequential. According to Evans (2003), these two systems "compete" for control over our judgments and actions, as people make decisions based either on past experience (based on what has worked in the past) or by constructing so-called mental models or by simulating future probabilities. Evans goes on to make the assumption that thinking in System 2 is both volitional and sensitive to verbal instruction, which is not true of System 1. Therefore, the effects of System 1 (in this case, belief bias) can only be counteracted indirectly, by asking participants to reason strictly deductively. One feature of System 2 can be thought of as its presumed ability to suppress or inhibit System 1's standard responses. However, it has a lower processing capacity, which requires more effort and the exclusion of attention to other things.

Sample of the text C

One model for investigating the human capacity to make decisions is the so-called dual-process models of thinking. According to Stanovich (2004), the experimental evidence from cognitive neuroscience and cognitive psychology agrees that brain functioning can be characterized by two distinct types of cognition. In doing so, each has distinct functions. In dual-process theories, processing in System 1 is characterized as automatic, based on heuristics, and relatively undemanding of computational capacity. Stanovich refers to this system in his terminology as a heuristic system and emphasizes that it combines the properties of automaticity, modularity, and heuristic processing. Automatic processing means that a certain process works, takes place, even when a person's attention is focused on another activity. Modular processing operates on the basis of separate, holistic knowledge. The heuristic search process is fast but risky; inaccuracies may result. Thus, the heuristic system answers quickly and automatically, but inference in it is biased due to the overall similarity of the stored stereotypes. System 2, which Stanovich (2004) refers to as analytic, combines various characteristics typical of volitionally controlled processing. Analytic cognitive processes are serial (in the case of the heuristic system they are parallel), rule-based, norm-based, computationally intensive, and are the focus of our awareness. In the case of an analytic system, the author is talking about problem solving at a conscious level. The analytical system is associated with individual differences in computational capacity, which is indirectly indicated by scores on intelligence tests and on tests of cognitive ability, and which is directly related to the level of working memory. An important function of the analytic system is that it serves as a tool capable of suppressing inappropriate responses generated by the heuristic system.

Appendix B

Examples of knowledge test items related to selected excerpts:

Q1: Explain in a few sentences the principle on which the "heuristic" System 1 works.

Q2: In a few sentences, describe what are the main advantages and what are the main disadvantages of the "heuristic" System1.

Q3: Explain in a few sentences the principle on which the "analytical " System 2 works.

Q4: In a few sentences, describe what are the main advantages and what are the main disadvantages of the "analytical" System 2.

Appendix C

Table showing the difficulty of question items.

Item	a	b1	b2
Q1	3.023	– 0.058	0.552

Item	a	b1	b2
Q2	3.370	– 0.309	0.309
Q3	2.549	– 0.223	0.460
Q4	2.724	– 0.180	0.575
Q5	2.653	– 0.563	0.585
Q6	2.119	– 0.046	0.516
Q7	1.301	0.376	1.305
Q8	1.377	0.376	1.087
Q9	1.138	– 0.222	1.541
Q10	1.250	0.199	0.869
Q11	1.255	0.559	0.834
Q12	1.236	0.074	0.655
Q13	1.474	0.499	0.893
Q14	1.598	0.048	0.423
Q15	1.143	0.993	1.537

A refers to the slope parameter. b1 and b2 refer to the two threshold parameters

Acknowledgements We thank Anna Pilatova, Ph.D., for proofreading and editing. Tomáš Prošek, Ph.D., for help with statistical analysis regarding the validity of our test. Lucia Vitekova, Ph.D., for providing financial support for this work through her grant from University of Ss. Cyril and Methodius [FPPV-63-2024].

Funding Open access funding provided by The Ministry of Education, Science, Research and Sport of the Slovak Republic in cooperation with Centre for Scientific and Technical Information of the Slovak Republic. This work was supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic [Grant No. VEGA 2/0026/21]. Grant recipient: Dr. Kamila Urban.

Data availability Data are available upon request from the first author.

Declarations

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

References

Alexander, P. A., & Judy, J. E. (1988). The interaction of domain-specific and strategic knowledge in academic performance. *Review of Educational Research, 58*(4), 375–404.

Baars, M., Khare, S., & Ridderstap, L. (2022). Exploring students' use of a mobile application to support their self-regulated learning processes. *Frontiers in Psychology., 13*, 793002. https://doi.org/10.3389/fpsyg.2022.793002

Barzilai, S., & Eseth-Alkalai, Y. (2015). The role of epistemic perspectives in comprehension of multiple author viewpoints. *Learning and Instruction, 36*, 86–103.

Barzilai, S., Zohar, A. R., & Mor-Hagani, S. (2018). Promoting integration of multiple texts: A review of instructional approaches and practices. *Educational Psychology Review, 30*(3), 973–999. https://doi.org/10.1007/s10648-018-9436-8

Bjørk, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). The MIT Press.

Bjørk, E. L., & Bjørk, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *& FABBS Foundation, Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Bjørk, R., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review Psychology, 64*, 417–444.

Braasch, J. L., McCabe, R. M., & Daniel, F. (2016). Content integration across multiple documents reduces memory for sources. *Reading and Writing, 29*, 1571–1598. https://doi.org/10.1007/s11145-015-9609-5

Bråten, I., Braasch, J. L. G., & Salmerón, L. (2020). *Reading multiple and non-traditional texts: New opportunities and new challenges.* In E. B. Moje, P. Afflerbach, P. Enciso, & N. K. Lesaux (Eds.), *Handbook of Reading Research* (Vol. V, pp. 79–98). Routledge.

Bråten, I., & Strømsø, H. I. (2011). Measuring strategic processing when students read multiple texts. *Metacognition and Learning, 6*(2), 111–130. https://doi.org/10.1007/s11409-011-9075-7

Britt, M. A., & Aglinskas, C. (2002). Improving students' ability to identify and use source information. *Cognition and Instruction, 20*, 485–522.

Britt, M. A., Perfetti, C. A., Sandak, R., & Rouet, J. F. (1999). Content integration and source Separation in learning from multiple texts. In S. R. Goldman, A. C. Graesser, & P. van den Broek (Eds.), *Narrative, comprehension, causality, and coherence: Essays in honor of Tom Trabasso* (pp. 209–233). Erlbaum.

Britt, M. A., & Rouet, J. F. (2012). Learning with multiple documents: Component skills and their acquisition. In J. R. Kirby & M. J. Lawson (Eds.), *Enhancing the quality of learning: Dispositions, instruction, and learning processes* (pp. 276–314). Cambridge University Press.

Britt, M. A., & Rouet, J. (2020). Multiple document comprehension. In: *Oxford research encyclopedia of education*.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1118–1133.

Callender, A. A., & McDaniel, M. A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology, 34*, 30–41.

Carrier, L. M. (2003). College students' choices of study strategies. *Perceptual & Motor Skills, 96*, 54–56.

Castells, N., Minguela, M., Solé, I., Miras, M., Nadal, E., & Rijlaarsdam, G. (2022). Improving questioning–answering strategies in learning from multiple complementary texts: An intervention study. *Reading Research Quarterly, 57*(3), 879–912.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Champely, S. (2020). pwr: Basic Functions for Power Analysis. R package version 1.3-0. https://CRAN.R-project.org/package=pwr

Dignath, C., Buettner, G., & Langfeldt, H.-P. (2008). How can primary school students learn self-regulated learning strategies most effectively? A meta-analysis on self-regulation training programmes. *Educational Research Review, 3*, 101–129.

Dunlosky, J., & Ariel, R. (2011). Self-regulated learning and the allocation of study time. *Psychology of Learning and Motivation, 54*, 103–140.

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

Dunlosky, J., & Rawson, K. A. (2005). Why does rereading improve metacomprehension accuracy? Evaluating the levels-of-disruption hypothesis for the rereading effect. *Discourse Processes, 40*, 37–56.

Dunlosky, J., & Rawson, K. (2012). Overconfidence produces underachievement: Inaccurate self-evaluations undermine students' learning and retention. *Learning and Instruction, 22*, 271–280.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising direction from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction, 24*(1), 58–61.

*Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454–459. https://doi.org/10.1016/j.tics.2003.08.012

Ferguson, L. E., Bråten, I., Strømsø, H. I., & Anmarkrud, Ø. (2013). Epistemic beliefs and comprehension in the context of reading multiple documents: Examining the role of conflict. *International Journal of Educational Research, 62*, 100–114.

Field, A. (2005). *Discovering Statistics using SPSS*. Sage Publications.

Fisher, J. S., & Radvansky, G. A. (2018). Patterns of forgetting. *Journal of Memory and Language, 102*, 130–141. https://doi.org/10.1016/j.jml.2018.05.008

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Fox, J., & Weisberg, S (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Fritz, C. O., Morris, P. E., Bjork, R. A., Gelman, R., & Wickens, T. D. (2000). When further learning fails: Stability and change following repeated presentation of text. *British Journal of Psychology, 91*, 493–511.

Gaelic, M. (2006). Dropouts on the web: Effects of interest and burden experienced during an online survey. *Journal of Official Statistics, 22*(2), 313–328.

Glover, J. A., & Corkill, A. J. (1987). Influence of paraphrased repetitions on the spacing effect. *Journal of Educational Psychology, 79*, 198–199.

Goldman, S.R. (2004). Cognitive aspects of constructing meaning through and across multiple texts. In N. Shuart-Faris & D. Bloome (Eds.), *Uses of intertextuality in classroom and educational research* (pp. 317–351). Information Age.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review, 19*, 126–134.

*Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux.

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17*, 471–479.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49*(4), 294–303. https://doi.org/10.1037/0003-066X.49.4.294

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

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

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.

Koriat, A. (2008). Easy comes, easy goes? The link between learning and remembering and its exploitation in metacognition. *Memory & Cognition, 36*(2), 416–428. https://doi.org/10.3758/MC.36.2.416

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 187–194.

Krug, D., Davis, T. B., & Glover, J. A. (1990). Massed versus distributed repeated reading: A case of forgetting helping recall? *Journal of Educational Psychology, 82*, 366–371.

Lee, H. Y., & List, A. (2023). Examining students' help-seeking when learning from multiple texts. *Contemporary Educational Psychology, 75*, 102232.

Lenth, R.V. (2021). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.6.0. https://CRAN.R-project.org/package=emmeans

List, A., Du, H., & Lee, H. Y. (2021). How do students integrate multiple texts? An investigation of top-down processing. *European Journal of Psychology of Education, 36*, 599–626.

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition, 1*, 18–26.

McNamara, D. S., Jacovina, M. E., and Allen, L. K. (2015). "Higher order thinking in comprehension," in Handbook of Individual Differences in Reading: Text and Context, ed. P. Afflerbach (Abingdon: Routledge), 164–176

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437–455.

Metcalfe, J. (2009). Metacognitive judgements and control of study. *Current Directory Psychological Science, 18*, 159–163.

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

Miele, D. B., Finn, B., & Molden, D. C. (2011). Does easily learned mean easily remembered? It depends on your beliefs about intelligence. *Psychological Science, 22*, 320–324.

Millis, K. K., Simon, S., & ten Broek, N. S. (1998). Resource allocation during the rereading of scientific texts. *Memory & Cognition, 26*, 232–246.
Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science, 5*, 207–213.
Nietfeld, J. L., & Schraw, G. (2002). The effect of knowledge and strategy training on monitoring accuracy. *The Journal of Educational Research, 95*, 131–142.
Nokes, J. D., Dole, J. A., & Hacker, D. J. (2007). Teaching high school students to use heuristics while reading historical texts. *Journal of Educational Psychology, 99*, 492–504.
OECD (2016). *Education at a Glance 2016: OECD Indicators*, OECD Publishing, Paris. https://doi.org/10.1787/eag-2016-en
Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences, 12*(6), 237–241. https://doi.org/10.1016/j.tics.2008.02.014
Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*, 422. https://doi.org/10.3389/fpsyg.2017.00422
Perfetti, C. A., Rouet, J. F., & Britt, M. A. (1999). Toward a theory of documents representation. In H. Van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representation during reading* (pp. 99–122). Erlbaum.
Pressley, M., & Woloshyn, V. (1995). *Cognitive strategy instruction that really improves children's academic performance*. Brookline Books.
Primor, L., & Katzir, T. (2018). Measuring multiple text integration: A review. *Frontiers in Psychology, 9*, 2294. https://doi.org/10.3389/fpsyg.2018.02294
Quinn, G. P., & Keough, M. J. (2002). *Experimental Designs and Data Analysis for Biologists*. Cambridge University Press.
R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/
R Studio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/
Radvansky, G. A., Zwaan, R. A., Curiel, J. M., & Copeland, D. E. (2001). Situation models and aging. *Psychology and Aging, 16*, 145–160.
Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning of key concepts in textbook materials. *European Journal of Cognitive Psychology, 19*(4–5), 559–579. https://doi.org/10.1080/09541440701326022
Rawson, K. A., Dunlosky, J., & Theide, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition, 28*(6), 1004–1010. https://doi.org/10.3758/BF03209348
Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology, 97*(1), 70–80. https://doi.org/10.1037/0022-0663.97.1.70
Rhodes, M., & Castel, A. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*, 615–625.
Richter, T., & Maier, J. (2017). Comprehension of multiple documents with conflicting information: A two-step model of validation. *Educational Psychologist, 52*(3), 148–166.
Rouet, J. F. (2006). *The skills of document use: From text comprehension to web-based learning*. Lawrence Erlbaum Associates.
Rouet, J., & Britt, M. A. (2011). Relevance processes in multiple document comprehension. In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 19–52). Information Age.
Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432.
Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*, 33–45. https://doi.org/10.1007/s11409-008-9031-3
Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting Self-Regulation in Science Education: Metacognition as Part of a Broader Perspective on Learning. *Research in Science Education, 36*, 111–139. https://doi.org/10.1007/s11165-005-3917-8
*Stanovich, K. E. (2004). *The robot's rebellion: Finding meaning in the age of Darwin.* University of Chicago Press.
Stine-Morrow, E. A. L., Gagne, D. D., Morrow, D. G., & DeWall, B. H. (2004). Age differences in rereading. *Memory & Cognition, 32*, 696–710.
Strømsø, H. I., Bråten, I., Britt, M. A., & Ferguson, L. E. (2013). Spontaneous sourcing among students reading multiple documents. *Cognition and Instruction, 31*(2), 176–203.

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. M. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes, 47*, 331–362.

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

*Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

*Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task, *Quarterly Journal of Experimental Psychology*, 12(3), 129–140, https://doi.org/10.1080/17470216008416717

Weinstein, Y., McDermott, K. B., Roediger, H. L., & III. (2010). A comparison of study strategies for passages: Rereading, answering questions, and generating questions. *Journal of Experimental Psychology: Applied, 16*, 308–316.

Wickham, H., François, R., Henry, L., Müller, K., (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

Wiley, J., Griffin, T. D., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *The Journal of General Psychology, 132*, 408–428.

Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology, 91*, 301–311.

Winne, P. H. (2015). Self-Regulated Learning. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences, 2nd Edition* (pp. 535–540). Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.25091-5

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist, 25*(1), 3–17. https://doi.org/10.1207/s15326985ep2501_2

*Concepts and theories in the texts for studying purposes are from these sources and publications.

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