How repeated studying and testing affects multimedia learning: Evidence for adaptation to task demands

Alexander Eitel \(^{a, b, *}\)

\(^{a}\) Leibniz-Institut für Wissensmedien, Tübingen, Germany  
\(^{b}\) University of Freiburg, Freiburg, Germany

**Abstract**

Two biases can occur in multimedia learning: overconfidence and over-reliance on text processing. The present research sought to identify these biases and to investigate whether they can be reduced, and hence learning fostered, when studying and testing are repeated. In 2 experiments (Exp.1: \(N = 79\), Exp.2: \(N = 52\)), students learned either with text only or with text and pictures (multimedia) about how the toilet flush works, gave judgments-of-learning (JOLs), were tested on the learning contents; afterwards this study-test cycle was repeated. Results from both experiments revealed stronger overconfidence due to multimedia in both study-test cycles (JOLs higher than learning outcomes). Eye movement data showed a relative increase in attention on the picture versus text from cycle 1 to cycle 2; this relative increase in attention was related to better learning outcomes. Repeated studying and testing thus helped to reduce over-reliance on text processing in multimedia learning, fostering performance.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

It is a well-established finding that when students learn with text and corresponding pictures, they achieve higher levels of recall and understanding than when they learn with text alone (i.e., multimedia effect; Butcher, 2014). This multimedia effect was mostly obtained by having students learn the instructional materials once, followed by the knowledge tests. Even though this setup is methodologically sound, it might be too constrained to capture the wealth of processes that can occur when learning with multimedia. For instance, when students prepare for an exam, they may repeatedly study the contents of the multimedia lesson and take practice tests between periods of study. How does such a behavior affect the multimedia effect? And which role do cognitive and metacognitive processes play in this? These questions were investigated in the present research; among others, by using eye tracking as a process measure. To this end, the present research extended the usual setup to study multimedia effects by repeating the study-test cycle and by investigating the effects on a process-level. Referring to theoretical frameworks on multimedia learning and metacognitive monitoring and regulation, it was tested whether repeated studying and testing would reduce biases in multimedia learning, thereby fostering study success.

1.1. Multimedia effect

Influential theories of multimedia learning (e.g., Mayer, 2014) mostly rely on cognitive processes to explain multimedia effects. According to the cognitive theory of multimedia learning (CTML; Mayer, 2014) learning with multimedia (i.e., text and pictures) entails selecting and organizing information from both text and pictures into two separate mental models in working memory. These two mental models are subsequently integrated with the help of prior knowledge. Referring to dual coding theory (Paivio, 1986), CTML assumes that when learning with text and pictures information will be stored in both a verbal and a nonverbal (pictorial) code. When trying to retrieve the processed information at a later point in time, this is possible by retrieving either the verbal or the pictorial or the integrated code. After learning with text (or picture) only, information can be retrieved only by one code in memory, which makes successful retrieval less likely. Therefore, better recall is assumed to result from learning with text and pictures compared to learning with text only, which has been confirmed by a vast number of empirical studies (see Butcher, 2014 for a review).
According to the integrative model of text and picture comprehension (ITPC; Schnotz, 2014), presenting a picture in addition to text facilitates mental model construction. Against the backdrop of the construction-integration model (Van Dijk & Kintsch, 1983), ITPC assumes that deeper comprehension is achieved only when readers construct a representation of the meaning of a message (i.e., text base) that can then be integrated into a mental model of the described situation with the help of prior knowledge. According to ITPC, presenting a picture in addition to text can facilitate construction of a richer mental model, because the picture as external representation shares a more direct relationship with the respective mental model than text does. Ideally, spatial relations from the picture can be directly mapped onto conceptual relations in the mental model (analogous structure mapping), so that subsequent mental model construction from text can be based on spatial relations initially extracted from the picture. The picture thus acts as a mental scaffold facilitating subsequent processing and comprehension of text (Eitel, Scheiter, Schüler, Nystöm, & Holmqvist, 2013). Therefore, presenting pictures in addition to text can foster comprehension, especially if pictures are processed early, and thus, prior to or simultaneously with the text (Eitel, Scheiter, & Schüler, 2013; Eitel & Scheiter, 2015).

Moreover, students' eye fixations fell on the picture as external representation shares a more direct relationship with the respective mental model than text does. Ideally, spatial relations from the picture can be directly mapped onto conceptual relations in the mental model (analogous structure mapping), so that subsequent mental model construction from text can be based on spatial relations initially extracted from the picture. The picture thus acts as a mental scaffold facilitating subsequent processing and comprehension of text (Eitel, Scheiter, Schüler, Nystöm, & Holmqvist, 2013). Therefore, presenting pictures in addition to text can foster comprehension, especially if pictures are processed early, and thus, prior to or simultaneously with the text (Eitel, Scheiter, & Schüler, 2013; Eitel & Scheiter, 2015).

Among the most famous multimedia heuristics are the anchoring and adjustment heuristic (Zhao & Linderholm, 2008) and the illusion of understanding heuristic (Johnson & Mayer, 2012; Mason, Pluchino, & Tornatora, 2013; Mason, Tornatora, & Pluchino, 2013). The illusion of understanding heuristic suggests that when learners perceive comprehension processes as being easier when these processes are based on passive perception of illustrations rather than on active elaboration of a text (illusion of understanding; Betrancourt, 2005; Küh, Scheiter, Gerjets, & Gambala, 2011; Paik & Schraw, 2013; Salomon, 1984). Such overconfidence can be detrimental to learning (Dunlosky & Rawson, 2012), because students may prematurely believe that they have reached an appropriate understanding of the multimedia contents. In consequence, they may invest too little effort (Paik & Schraw, 2013) and allocate study time ineffectively by premature study termination.

To conclude, there are two biases in multimedia learning. First, because students lack knowledge about how to prepare for an upcoming knowledge test, they can be prone to over-rely on the processing of text. Second, due to the multimedia heuristic students can become overconfident, being related with reduced study time and effort in multimedia learning. One potential means to diminish such biases is to give students the opportunity to repeat studying after assessing their understanding in a knowledge test.

1.3. Reducing biases: repeated studying and testing

In more basic memory research, the study of learning often involves several cycles of learning and testing (i.e., multitrial learning), sometimes as many as it takes to recall the information reliably (Karpicke & Roediger, 2008; Thiede & Dunlosky, 1999).
Unsurprisingly, in this research participants usually show improvements in recall performance that occur with repeated studying and testing. There are several reasons for improvements in recall due to repeated studying and testing, on which I will elaborate in the following.

First, participants benefit from being exposed to the same information twice, or multiple times, because rehearsing information, even in the same format, typically makes it easier to remember it (Mayer & Johnson, 2008; Ortegaen, Serra, & Englund, 2014). However, these knowledge gains are sustainable only if not just studying, but also testing is repeated. Repeated testing has been found to be much more beneficial to long-term retention than repeated studying (i.e., testing effect), because retrieval — as it will be required in the final test — is being practiced, and because testing typically leads to more elaboration of recalled information, thus producing stronger memory traces (Karpicke & Roediger, 2008; Roediger & Butler, 2011). Repeated testing and testing is a particular instance of the testing effect; that is, an indirect testing effect or test-potentiated learning (Arnold & McDermott, 2013).

Test-potentiated learning means that in addition to fostering retention of previously recalled information (direct testing effect), testing enhances subsequent learning of information that learners previously failed to recall (Arnold & McDermott, 2013). Thus, repeated studying of the same information can be adapted to the demands of the preceding test. If also testing is repeated, students know what to expect from the next test so that they can monitor and control their learning accordingly (transfer-appropriate processing; Morris, Bransford, & Franks, 1977), resulting in increased metacognitive monitoring accuracy and test performance (Arnold & McDermott, 2013; Dunlosky, Rawson, & Middleton, 2005). In line with the idea of transfer-appropriate processing (Morris et al., 1977), Thiede, Wiley, and Griffin (2011) found that test performance and monitoring accuracy were superior when students received the kind of test they expected as opposed to an unexpected test. Hence, learning activities can be better regulated the more students are aware of the goals and demands of studying. Following Metcalfe (2002), depending on the goals and expectations, students adapt their study behavior by selectively allocating study time to items that maximize learning outcomes (i.e., region of proximal learning). Similarly, students adapt their processing behavior to the complexity of the task demands (Pieschl, Stahl, Murray, & Bromme, 2012). Specifically, in the study of Pieschl et al. learners showed superficial strategies for simple tasks, whereas they made more use of deep elaborative strategies for complex tasks. Using deep elaborative strategies for complex tasks was predictive for better performance on this task. Thus, repeated studying and testing can benefit recall and comprehension, because learners can adapt their processing to the task demands after receiving the first test. In contrast to previous research, the present research made use of eye tracking to identify adaptation to task demands on a fine-grained processing level.

Second, repeated studying and testing may foster learning, because overconfidence may be reduced. Specifically, after being tested for recall and comprehension of learning materials, initial overconfidence can be reduced, because success in retrieving information in the first test provides students with a valid cue about how well they learned with the given materials (Finn & Metcalfe, 2014). Confidence levels are hence better matched to the actual learning success in a subsequent study-test cycle, as already shown by studies on vocabulary learning (i.e., better calibration scores in tests 2 and 3 than in test 1; Carpenter & O’Dvorak, 2012). Accordingly, success in retrieving information in the first test has been shown to predict confidence levels when restudying the materials (memory for past test heuristic; Finn & Metcalfe, 2014). Materials that initially seemed easy to be learned might receive more deliberate attention after recognizing memory gaps or comprehension deficits in a first test. In consequence, overconfidence might even shift towards underconfidence in subsequent trials (‘overconfidence with practice effect’; Koriat, Sheffer, & Ma’ayan, 2002). This effect can be desirable to learning, because reducing initial over-confidence might increase study effort and effective study time allocation, thereby increasing performance in the final test.

1.4. Reducing biases in multimedia learning

Over-reliance on text processing in multimedia learning may be reduced by repeating the study-test cycle, because repeated studying of the same text and pictures can be adapted to the demands of the preceding test (adaptation to task demands; Morris et al., 1977; Pieschl et al., 2012). In particular, because after the initial round of studying and testing students know what to expect from the next test, they should use both text and pictures more deliberately to extract specific information needed for the test in cycle 2; hence, they may not show the bias of over-reliance on the text anymore, and attend relatively more to the picture. In eye tracking terms, more and longer eye fixations should be directed towards the picture as a measure of cognitive processing depth (cf. Rayner, Chace, Slattery, & Ashby, 2006). This should in turn be beneficial for learning outcomes. Without the picture, students may adapt their learning to task demands by spending more time carefully re-reading text according to their knowledge deficits, and by investing more effort into mentally imagining the learning contents (i.e., longer fixation durations), which should foster learning outcomes. In general, students should study longer in cycle 2, the more difficulties they faced solving test 1, reflecting adaptation to task demands.

More pronounced overconfidence due to multimedia (cf. multimedia heuristic; Serra & Dunlosky, 2010) could be reduced when studying and testing are repeated, because success in retrieving information in the first test would be used as a valid cue to judge performance in test 2 (underconfidence with practice; Koriat et al., 2002). Hence, students who learn with multimedia could profit even more from repeating studying and testing than students who learn with text, resulting in better regulation and a stronger multimedia effect in study-test cycle 2.

1.5. Present research and hypotheses

The main focus of the present research was to investigate how the multimedia effect develops across study-test cycles and which (meta-)cognitive processes are associated with it. Therefore, in two experiments, instructional materials were either presented by text only or by text with pictures (multimedia), and both the instructional materials and the knowledge tests were presented to students twice in succession. I assessed students’ monitoring and regulation processes in the first and second study-test cycle on different levels, namely by assessing JOLs, learning outcomes, calibration scores (JOLs — learning outcomes), study times, and eye movements (in Experiment 2).

Using this experimental design, several effects were presupposed to occur the way they occurred in prior research. In particular, it was presupposed that learning with text and pictures yields higher JOLs (multimedia heuristic; Serra & Dunlosky, 2010) and better learning outcomes than learning with text only (multimedia effect; Butcher, 2014). Moreover, given repeated exposure to the same materials and tests, study times should be reduced (McCrueen, Magliano, & Schraw, 2006), and learning outcomes should be better in the second than in the first study-test cycle (rehearsing and test-potentiated learning; Arnold & McDermott, 2013; Mayer & Johnson, 2008).
Adaptation to task demands:
1) In both conditions, the lower the performance in test 1, the longer study times should be in cycle 2 (Hypothesis 1).
2) In the multimedia condition, the decrease in study times from cycle 1 to cycle 2 should be associated with a steeper decrease in attention on the text than on the picture (Hypothesis 2a). Mean fixation durations on the picture as a measure of cognitive effort should be longer in cycle 2 than in cycle 1 (Hypothesis 2b). These measures should be related to better study outcomes (Hypothesis 2c).
3) In the text-only condition, students should attend longer to text (Hypothesis 3a) and have longer mean fixation durations as a measure of reading effort (Hypothesis 3b) compared to students in the multimedia condition, especially in study-test cycle 2. These measures should be related to better learning outcomes (Hypothesis 3c).

Underconfidence with practice:
1) In both conditions, students should be overconfident in study-test cycle 1 (JOLs higher than learning outcomes), which should be especially pronounced with multimedia (Hypothesis 4a). Overconfidence should be reduced after the first test, reflected in calibration scores near zero in both conditions in cycle 2 (Hypothesis 4b).
2) Greater confidence reduction in the multimedia than in the text-only condition should be associated with longer study times (Hypothesis 5a) and greater learning gains (Hypothesis 5b) in cycle 2 than in cycle 1 (i.e., stronger multimedia effect in cycle 2).
3) In both conditions lower performance in test 1 should lead to stronger confidence reduction across cycles, being related with a smaller decrease in study times from cycle 1 to cycle 2 (Hypothesis 6).

2. Experiment 1

2.1. Method

2.1.1. Participants and design
Seventy-nine undergraduate students (64 women, 15 men, $M_{age} = 23.14$ years, age range: 18—35 years) from a German University took part in the experiment for payment (12 Euro). Students were recruited from a student database, for which they volunteered to being listed. Students with study majors in physics or engineering were not allowed to take part in the experiment. Participants were randomly assigned to two experimental groups — multimedia (text and picture) vs. text only — so that there were 40 participants in the multimedia and 39 participants in the text only group.

2.1.2. Materials
The instructional materials comprised seven pages with expository text (and pictures) about the structure and the functioning of the toilet flushing system (Mayer, Hegarty, Mayer, & Campbell, 2005; Paik & Schraw, 2013). On the first two pages of the instruction, the relevant components of the toilet flushing system and how they were spatially related to each other were introduced (see Fig. 1). On pages 3 to 7, the mechanism underlying the toilet flush was explained by describing the behavior of the system’s components at each step of the flushing process, starting with “when the handle is pressed down, the connecting rod is pulled up etc.”. In total, the German version of the text comprised 497 words. Text was longer on the first two pages (135.5 words on average per page) than on the following pages (45.2 words on average per page). In the multimedia condition, a picture was presented along with the text on each page (see Fig. 1, left column). The picture was always displayed centrally on the screen with associated text beneath it. Pictures were schematic but colored line drawings about the structure and components of the toilet tank; they represented information from the text (representation function; Carney & Levin, 2002). Pictures were very similar across the seven pages of the instruction — only the state of the toilet system was changed across pages (in order of the causal chain). Moreover, on the first two pages, the crucial components of the toilet flushing system were labeled in the pictures. The labels were printed in boldface in the text to enhance their salience, so that also students in the text-only condition knew that those components were the most important information on that page (see Fig. 1, right column). In the text-only condition, text was always presented in the same location as in the multimedia condition. The text was identical in the multimedia and in the text-only condition. Since the text provided a detailed description of the structure and functioning of the toilet flushing system, studying text only was sufficient to correctly answer the questions of the knowledge test.

2.1.3. Measures
To control for students’ abilities that are relevant with respect to the dependent variables in the experiments, questions concerning students’ latest school grade in physics and students’ estimated prior knowledge for the toilet flushing system as well as a shortened paper-pencil version of the paper folding test (PFT; Ekstrom, French, & Harman, 1976) were administered. The PFT assesses students’ spatial reasoning abilities. Study times were assessed via computer log files for each of the seven pages of the instruction in both cycles. Judgments of learning (JOLs) were assessed with pen on paper. JOLs asked students to rate on a scale from 0 to 100 “how confident are you that you will correctly answer questions based on the information just learned?” (Serra & Dunlosky, 2010). Moreover, students’ judgments of recall (“how confident are you that you will correctly recall information just learned”) and comprehension (“how confident are you that you will be able to explain why a toilet flush does not work the way it should, based on the information just learned”) were assessed. However, high correlations between judgments of recall and judgments of comprehension in both experiments (.80 < r < .87) suggest that students did not distinguish between their ability to recall and to understand, which is why only the overall judgment of learning (JOL) was used for further analysis.

To assess students’ learning outcomes, questions addressed recall, troubleshooting, and pictorial matching of information from text (and pictures). All responses were scored by at least one rater who was always blind to the experimental condition. Questions of recall and troubleshooting were in an open answer format (Hegarty, Križ, & Cate, 2003). There was one recall question: ‘Describe how a toilet tank works. Imagine that you push down on the handle of the toilet tank. Describe step-by-step what happens to each of the other parts of the tank as it flushes.’ There were four troubleshooting questions: (1) ‘Suppose you push down on the handle of the toilet tank but water does not flush into the toilet bowl. Explain all the possible things that could be wrong.’ (2)
Suppose that after flushing the toilet, you notice that water is continuously running into the toilet tank. Explain all the possible things that could be wrong. (3) ‘Suppose that after you flush the toilet, water continues to run into the toilet bowl without stopping. Explain all of the possible things that could be wrong.’ and (4) ‘What would happen if the float were to break off from the float arm? What would happen if the upper and lower disks were to stick to each other in the siphon bell? What stops the water from flushing out of the tank?’ In addition, there were three pictorial matching questions, in which participants were asked (1) to label the 10 major parts of the toilet tank, (2) to draw the connecting rod, the upper, and the lower disc, and (3) to draw the float, the float arm, the inlet valve, and the incoming water in a diagram of the toilet flush, in which those parts had been removed.

A pre-existing list of the 19 major units of the text was used for the scoring of the recall test (Hegarty et al., 2003; Exp. 1). Participants received one point for each major idea unit that they recalled, regardless of wording, leading to a maximum score of 19 points. Cronbach’s alpha for this question was .76. Similarly, for the troubleshooting questions, a list of correct answers and criteria for correctness was produced beforehand. A correct answer yielded one point. The final score of the troubleshooting test was determined by adding up all points for the troubleshooting questions (Mayer et al., 2005; Exp. 2), which yielded a maximum of 39 points. Cronbach’s alpha for this question was .69. Provided that I analyzed the overall judgment of learning (JOL) as the subjective measure of performance, I created the corresponding objective measure of performance by adding the individual scores for recall and troubleshooting per participant to result in a composite score. Justifications for this procedure can be provided on a conceptual and empirical level. First, the present hypotheses address learning outcomes in general, and do not distinguish between recall and troubleshooting. Second, there is a substantial overlap between the two: recall of knowledge is a necessary prerequisite to transfer knowledge to new situations (for troubleshooting). This is corroborated on an empirical level by substantial correlations between recall and troubleshooting performance in both experiments (.63 ≤ r ≤ .67) and by a Cronbach’s alpha that was higher for the composite score (.83) than for the separate scores of recall and troubleshooting.

A second rater scored 23% of the open questions assessing recall and troubleshooting performance (from 18 out of 79 participants), and Krippendorff’s alpha was calculated as the measure of inter-rater agreement (Hayes & Krippendorff, 2007). Inter-rater agreement for the composite score of recall and troubleshooting was very high (Krippendorff’s alpha = .96) so that only one rater continued scoring the rest of the data.

As in previous research (Finn & Metcalfe, 2014), scores for calibration accuracy were calculated by subtracting the scores for the composite score of recall and troubleshooting (standardized to scale from 0 to 100) from the JOLs (scale from 0 to 100). To standardize the composite scores, they were divided by the maximum number of points in the recall plus troubleshooting test (58), and then multiplied by 100, so that as for JOLs, composite scores were on a scale with a minimum value of 0 and a maximum of 100. Thus, a value of (or near) zero reflects accurate calibration, negative values reflect underconfidence, and positive values reflect overconfidence.

For the pictorial matching questions, a list of correct labels and drawings was produced beforehand. The overall score for pictorial matching was obtained by subsuming all points from all three questions (max. 20 points). Cronbach’s alpha for this score was .77.

### 2.1.4. Procedure

Participants were tested in groups of three to six persons per session. Participants were seated at desks in individual cubicles without visual contact to other participants. First, participants were required to give their informed consent and to answer demographic questions. Thereafter, written instructions and the instructional materials (‘how toilet flushing system works’) were presented on a separate laptop for each participant in a session (via e-prime® 2.0 professional). Participants were instructed that they could navigate forward (but not backward) through the pages at their own pace by pressing the space bar. Moreover, participants
were instructed to study the materials thoroughly, so that they would be able to later answer questions about their contents. On the first two pages of the instruction, the names of the toilet flushing system’s components were introduced, and participants were instructed to memorize these names because they would need them to learn about the system’s functioning in the following. After participants finished studying the seven pages of the instruction, they were required to judge how confident they were regarding their learning success (JOL rating), followed by the open questions assessing recall (6 min) and troubleshooting (11.5 min in total). Participants were given 2.5 min for each of the first three troubleshooting questions, and 4 min for the fourth troubleshooting question. They did not receive feedback about their performance in the tests. All participants of a group started responding to each question simultaneously. They were instructed to stop writing as soon as the time to respond to a question was up (signaled by an alarm clock). Participants who finished responding to a question early (i.e., before time was up) were not allowed to continue working on the next question; rather, they had to wait until all participants of the group were asked to start with the next question. Participants were not allowed to return to previous questions to review their previous answers.

After completing the recall and troubleshooting questions, participants started to work on the PFT for a maximum of three minutes. Afterwards, the identical procedure of presenting instructional materials, the JOL rating, and the posttest questions was repeated (see Fig. 2). Prior to studying the instructional materials for a second time, participants were told that they would receive the same posttest questions again. After responding to the posttest questions in the second study-test cycle, the pictorial matching questions were additionally presented. In total, students had 7 min to answer to these questions (2 × 2 min + 1 × 3 min). Pictorial matching questions were not presented in the first cycle, because presenting these pictures prior to second study-test phase would have acted as a feedback and might have influenced learning differently in the text only and the multimedia condition. After responding to the pictorial matching questions, students were debriefed and thanked for participation. The whole experiment lasted approximately 90 min.

### Results

Results were analyzed in a three-step manner. First, students’ latest school grades in physics, their estimated prior knowledge and their spatial reasoning abilities were analyzed to determine whether the groups were comparable regarding these characteristics. Second, students’ JOLs, calibration scores, learning outcomes, and study times were analyzed by means of 2 × 2 ANOVAs. Third, relations between learning outcomes, JOLs and study times were analyzed. Partial eta-squared ($\eta^2_p$) and Cohen’s $d$ were reported as measures of effect size. For eta-squared, values of .01, .06, and .14 correspond to small, medium, and large effect sizes, whereas for Cohen’s $d$ values of .30, .50, and .80 are considered small, medium, and large effect sizes, respectively (Cohen, 1988). The reported $p$-values are estimations for the exact $p$-values in the population.

#### Group characteristics

Descriptive values are shown in Table 1. One-way ANOVAs revealed that students in the two experimental groups (multimedia vs. text only) did not differ concerning their latest school grade in physics, $F < 1$, estimated prior knowledge for the toilet flushing system, $F(1, 77) = 2.06, p > .05$, and spatial reasoning abilities, $F(1, 77) = 1.55, p > .05$. Thus, groups were comparable regarding their relevant characteristics for the study.

#### JOLs

Descriptive values are shown in Table 2. A 2 × 2 ANOVA with cycle (1 vs. 2) and experimental condition (multimedia vs. text only) as factors, and magnitude of the JOL rating as dependent variable was conducted. It revealed main effects of cycle, $F(1, 77) = 72.98, p < .001, \eta^2_p = .49$, and experimental condition, $F(1, 77) = 26.12, p < .001, \eta^2_p = .25$, but no interaction, $F < 1$. Judgment magnitude was greater in cycle 2 than in cycle 1, and it was consistently greater for the multimedia group than for the text-only group.

#### Calibration scores

Descriptive values can be found in Table 2. One-sample $t$-tests revealed that in both cycles and in both conditions, calibration

---

**Table 1**

<table>
<thead>
<tr>
<th>Demographic variable</th>
<th>Experiment 1</th>
<th>All subjects</th>
<th>Experiment 2</th>
<th>All subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Multimedia</td>
<td>Text only</td>
<td>Multimedia</td>
<td>Text only</td>
</tr>
<tr>
<td>Latest school grade in physics (min. – 1, max. – 6)</td>
<td>2.72 (.02)</td>
<td>2.84 (.02)</td>
<td>2.56 (.82)</td>
<td>2.89 (.09)</td>
</tr>
<tr>
<td>Estimated prior knowledge (min. – 1, max. – 5)</td>
<td>1.66 (.76)</td>
<td>1.97 (.11)</td>
<td>1.64 (.81)</td>
<td>1.54 (.69)</td>
</tr>
<tr>
<td>Spatial reasoning abilities (min. – 0, max. – 10)</td>
<td>6.15 (.21)</td>
<td>6.72 (.95)</td>
<td>6.64 (.21)</td>
<td>7.41 (.55)</td>
</tr>
</tbody>
</table>

---

**Fig. 2.** Time course of both experiments. In both cycles, participants received the same instructional materials, JOL ratings, recall and transfer questions.
scores were greater than zero (all \( p < .001 \)), thereby reflecting general overconfidence. The ANOVA with cycle (1 vs. 2), experimental condition (multimedia vs. text only), and scores for calibration as dependent variable revealed main effects of cycle \( F(1,77) = 16.00, p < .001, \eta^2_p = .17 \), and experimental condition, \( F(1,77) = 24.19, p < .001, \eta^2_p = .24 \), and again no interaction, \( F < 1 \). Overconfidence was not reduced but even increased from cycle 1 to cycle 2 in both conditions. Moreover, overconfidence was stronger for students learning from multimedia than from text only (see Fig. 3, upper-right panel). To shed more light onto the calibration scores on an individual level, scatterplots for JOLs versus learning outcomes are presented in Fig. 5 (top row). The overall positive relationships between JOLs and learning outcomes suggest that no Dunning-Kruger effect\(^1\) emerged.

2.2.4. Learning outcomes

Descriptive values are shown in Table 2. I conducted a \( 2 \times 2 \)-ANOVA with cycle as within-subjects factor (cycle 1 vs. cycle 2), experimental condition (multimedia vs. text only) as between-subjects factor, and the composite score of recall and troubleshooting as dependent variable. The analysis revealed significant main effects of cycle, \( F(1,77) = 40.73, p < .001, \eta^2_p = .35 \), and experimental condition, \( F(1,77) = 4.08, p = .047, \eta^2_p = .05 \), but no interaction, \( F(1,77) = 1.09, p > .05 \). Students in both groups had better learning outcomes in cycle 2 than in cycle 1, and students in the multimedia group consistently outperformed students in the text-only group (see Fig. 3, upper-left panel).

Furthermore, a t-test revealed that in the pictorial matching task, which had been assigned only in cycle 2, students in the multimedia condition had better scores than students in the text-only condition, \( t(77) = 5.24, p < .001, d = 1.18 \).

2.2.5. Study times

Descriptive values are shown in Table 2. The \( 2 \times 2 \) ANOVA with cycle as within-subjects factor (1 vs. 2), experimental condition (multimedia vs. text only) as between-subjects factor, and the summed study time across the seven pages of the instruction as dependent variable was conducted. It revealed a main effect of cycle, \( F(1,77) = 91.34, p < .001, \eta^2_p = .54 \), no main effect of experimental condition, \( F < 1 \), but a significant interaction,

### Table 2

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Multimedia</th>
<th>Text only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cycle 1</td>
<td>Cycle 2</td>
<td>Combined</td>
</tr>
<tr>
<td>JOLs (min. – 0, max. 100)</td>
<td>67.00 (14.80)</td>
<td>77.70 (14.36)</td>
</tr>
<tr>
<td>Calibration scores (min. – – 100, max. – 100)</td>
<td>40.58 (13.48)</td>
<td>45.80 (12.72)</td>
</tr>
<tr>
<td>Learning outcomes (recall + transfer; min. – 0, max. – 58)</td>
<td>15.33 (4.09)</td>
<td>18.50 (4.21)</td>
</tr>
<tr>
<td>Pictorial matching scores (min. – 0, max. – 20)</td>
<td>17.29 (4.73)</td>
<td>17.29 (4.73)</td>
</tr>
<tr>
<td>Study times (in seconds)</td>
<td>302.22 (79.75)</td>
<td>186.72 (48.54)</td>
</tr>
</tbody>
</table>

Fig. 3. Results for learning outcomes and calibration scores as a function of experimental condition (text only vs. multimedia) and cycle (1 vs. 2) in Experiment 1 and Experiment 2.

\(^1\) Dunning-Kruger effect means that relatively unskilled individuals estimate their ability as much higher than it actually is, whereas highly skilled individuals underestimate their relative competence (for further information, see Kruger & Dunning, 1999).
Post hoc comparisons showed that there were no significant differences in study times in the first cycle of learning with multimedia versus text only ($p > .05$). However, students studied for a shorter time with multimedia than with text only in the second cycle, ($p = .001, d = .79$), explaining the significant interaction.

2.2.6. Link between learning outcomes, JOLs, and study times

To test whether lower performance in test 1 would be associated with longer study times in cycle 2 compared to cycle 1, a linear regression analysis was conducted with performance in test 1 as predictor, and the study time difference between cycle 2 and cycle 1 as criterion. Results confirmed that lower performance in test 1 was associated with relatively higher study times in cycle 2 compared to cycle 1, $\beta = -.36, t(78) = -3.43, p = .001$. To test whether lower performance in test 1 was associated with longer study times, because of a stronger confidence reduction from cycle 1 to cycle 2, a simple mediation model was calculated (Preacher & Hayes, 2004). In this model, the study time difference between cycle 2 and cycle 1 was entered as criterion, performance in test 1 as predictor, and the

<table>
<thead>
<tr>
<th>Multimedia</th>
<th>Text only</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Graph 1]</td>
<td>![Graph 2]</td>
</tr>
</tbody>
</table>

Fig. 4. Total number of eye fixations on text and on picture in the multimedia condition, and on text and on the rest of the screen (where picture was located in multimedia condition) in the text-only condition across the two cycles of the instruction in Experiment 2.

<table>
<thead>
<tr>
<th>Cycle 1</th>
<th>Cycle 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Graph 3]</td>
<td>![Graph 4]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exp.1</th>
<th>Exp.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Graph 5]</td>
<td>![Graph 6]</td>
</tr>
</tbody>
</table>

Fig. 5. Scatterplots to visualize the relation between judged (JOL scores) and actual learning outcomes (Composite test score) across cycles and experiments.

$F(1,77) = 18.54, p < .001, \eta^2_p = .19$. Post hoc comparisons showed that there were no significant differences in study times in the first cycle of learning with multimedia versus text only ($p > .05$). However, students studied for a shorter time with multimedia than with text only in the second cycle, ($p = .001, d = .79$), explaining the significant interaction.
3.1.3. Measures

The first cycle's measures replicated Experiment 1 and included eye tracking. To analyze the relationship between performance in test 1 and study times in cycle 2 compared to cycle 1, we used eye tracking to observe students' attention to text and pictures. The measure for students' overall attention distribution on text and pictures was the fixation count (i.e., number of fixations) on the AOIs for text and pictures, with the hypothesis that a higher number of fixations reflects a higher degree of cognitive processing (eye-mind hypothesis; Just & Carpenter, 1980; Rayner et al., 2006). As a measure of the ease or difficulty associated with how students processed text, we analyzed the mean duration of a fixation on the text, since longer mean fixation durations in reading are known to reflect a higher effort to understand text (Rayner et al., 2006; Rayner & Duffy, 1986). Moreover, we calculated the overall fixation time (fixation count * mean duration of a fixation) to assess how long the materials were attended in both conditions. The analysis was based on fixation count/time instead of on dwell time (comprising fixations and saccades), because only fixations allow for information intake and conscious cognitive processing (Holmqvist et al., 2011). Hence, fixation count/time is a more precise measure of cognitive processing than dwell time. Fixations (and saccades) were detected with a saccade velocity detection algorithm. An eye fixation was recorded as long as the eyes were relatively motionless, that is, when they moved slower than 40° of visual angle per second.

Concerning learning outcomes measures, Cronbach's alpha for the recall question was 0.75, for the troubleshooting questions 0.52, and for the pictorial matching questions 0.78. Cronbach's alpha for all questions in an open format (recall and troubleshooting) was 0.78. A second rater scored 31% of the data (from 17 out of 55 participants). Inter-rater agreement was sufficiently high (Krippendorff's alpha = .86) so that only one rater continued scoring the remaining data.

3.1.4. Apparatus

E-prime® 2.0 professional was used as software for the experiment. Stimuli were presented on a monitor that had a 1280 x 1024 pixel-resolution, a size of 29.54 x 24.37 degrees of visual angle, and a refresh rate of 60 Hz. Eye movements were recorded by means of a video-based eye tracking device (iView X™ highspeed 1250) from SensoMotoric Instruments (SMI, Teltow, Germany) with a 500 Hz sampling rate. Fixations and saccades were detected with the saccade velocity detection algorithm using the default settings of BeGaze3.4™.

3.1.5. Procedure

Procedure was the same as in Experiment 1 (see Fig. 2), with the exception that the eye tracking system had to be calibrated to the participants' gaze prior to presenting the instructional materials in both study-test cycles. This was done using a manually controlled, 13-point calibration procedure. Calibration of the eye tracking system was considered successful when validation accuracy was better or equal to a mean deviation of 1 degree of visual angle in both the vertical and horizontal direction.

3.2. Results

Results were analyzed as in Experiment 1, with the exception of the eye movement analyses including their relations to performance.

3.2.1. Group characteristics

Descriptive values are shown in Table 1. One-way ANOVAs revealed that students in the two experimental groups (multimedia vs. text only) did not differ significantly with regard to their latest school grade in physics, F(1, 50) = 1.50, p > .05, estimated prior knowledge for the toilet flushing system, F(1, 50) < 1, or spatial reasoning abilities, F(1, 50) = 2.25, p > .05. The groups were thus comparable...
3.2.2. JOLs

Descriptive values are shown in Table 3. A 2 × 2 ANOVA with cycle (1 vs. 2), and condition (multimedia vs. text only) as factors, and JOLs as dependent variable was conducted. It revealed main effects of cycle, \(F(1,50) = 20.92, p < .001, \eta^2_p = .30\), and condition, \(F(1,50) = 9.02, p = .004, \eta^2_p = .15\), and no interaction, \(F < 1\). As in Experiment 1, judgment magnitude was greater in cycle 2 than in cycle 1, and it was greater for the multimedia group than for the text-only group.

3.2.3. Calibration scores

Descriptive values can be found in Table 3. One-sample t-tests for calibration scores revealed that in both cycles and in both conditions, values were consistently greater than zero (all \(p_s < .001\), again reflecting a general overconfidence. The ANOVA with cycle (1 vs. 2) and condition (multimedia vs. text only) as factors, and calibration accuracy as dependent variable revealed no main effect of cycle \(F < 1\), a main effect of condition, \(F(1,50) = 6.37, p = .02, \eta^2_p = .11\), and no interaction, \(F < 1\). Thus, as in Experiment 1, there was a larger overconfidence due to multimedia that was not reduced by repeated studying and testing. In contrast to Experiment 1, calibration scores did not increase from cycle 1 to cycle 2 (see Fig. 3, lower-right panel). Scatterplots for JOLs versus learning outcomes again showed no Dunning-Kruger effect (see Fig. 5, bottom row).

3.2.4. Learning outcomes

Descriptive values are shown in Table 3. The 2 × 2-ANOVA with cycle (1 vs. 2) and experimental condition (multimedia vs. text only) as factors, and learning outcomes as dependent variable, revealed main effects of cycle, \(F(1,50) = 77.16, p < .001, \eta^2_p = .61\), and experimental condition, \(F(1,50) = 3.99, p = .05, \eta^2_p = .07\), and no interaction, \(F < 1\). Thus, as in Experiment 1, learning outcomes increased similarly in both groups from cycle 1 to cycle 2, and they were consistently higher in the multimedia than in the text-only group, hence showing a robust multimedia effect (see Fig. 3, lower-left panel).

Furthermore, a t-test revealed that students in the multimedia condition had better scores for pictorial matching than students in the text-only condition, \(t(43.39) = 5.22, p < .001, d = 1.42\).

3.2.5. Study times

Descriptive values are shown in Table 3. The 2 × 2 ANOVA with cycle (1 vs. 2) and condition (multimedia vs. text only) as factors, and overall study time as dependent variable revealed a main effect of cycle, \(F(1,50) = 22.53, p < .001, \eta^2_p = .31\), no main of experimental condition, \(F < 1\), and no significant interaction, \(F(1,50) = 1.21, p > .05\). Thus, in contrast to Experiment 1, students had comparable study times in both conditions. Those decreased in a parallel fashion from cycle 1 to cycle 2.

3.2.6. Link between learning outcomes, study times, and JOLs

A linear regression analysis was conducted with performance in test 1 as predictor, and the study time difference between cycle 2 and cycle 1 as criterion to test whether lower performance in test 1 led to relatively longer study times in cycle 2. As expected, and in line with Experiment 1, lower performance in test 1 was associated with relatively higher study times in cycle 2, \(\beta = .29, t(51) = -2.13, p = .04\). To test whether the link between performance in test 1 and study time differences could be explained by changes in confidence levels, a simple mediation model was calculated (Preacher & Hayes, 2004). In this model, the study time difference between cycle 2 and cycle 1 was entered as criterion, performance in test 1 as predictor, and the JOL difference between cycle 2 and cycle 1 as the proposed mediator. Results of the modeling did not reveal that changes in confidence levels mediated the relationship between performance in test 1 and study times in cycle 2 compared to cycle 1, \(\beta = -1.56, p > .05, 95\% CI [−.09, .01]\).

3.2.7. Eye movements

Eye movement data of six persons had to be excluded from the analysis due to severe problems with the eye tracking recording (i.e., many missing data points, strong drift or very low accuracy in calibrating the system). Accordingly, eye movement data from 47 participants (34 women, 13 men, \(M_{age} = 21.70\), age range: 18–29 years) were analyzed. Descriptive values are shown in Table 4.

With respect to the amount of attention spent on text and pictures in the multimedia condition (measured via total fixation count on text-AOI vs. picture-AOI), a 2 × 2 repeated-measures ANOVA with cycle (1 vs. 2) and medium (text vs. picture) as factors revealed a significant interaction, \(F(1,22) = 22.30, p < .001, \eta^2_p = .58\), and medium, \(F(1,22) = 65.05, p < .001, \eta^2_p = .75\), as well as an interaction, \(F(1,22) = 9.93, p = .01, \eta^2_p = .31\). Post hoc comparisons demonstrated that although both text (\(p < .001, d = 1.17\)) and pictures (\(p = .01, d = .65\)) received less attention in cycle 2 than in cycle 1, the decrease was steeper for the text than for the picture, explaining the significant interaction (see Fig. 4). Hence, a more balanced attention distribution between text and picture was found in cycle 2 than in cycle 1. Moreover, as expected mean fixation durations on the picture increased from cycle 1 to cycle 2 in the multimedia condition, \(F(1,22) = 4.45, p = .046, \eta^2_p = .17\).

To test whether a more balanced attention distribution and/or whether longer mean fixation durations were, in turn, related to better study outcomes, I calculated the relative score for attention on the picture versus text in cycle 2 (fixation count on picture – fixation count on text), and standardized this value by subtracting it from the relative score for attention on the picture versus text in cycle 1 (count difference: \(M = 23.28, SD = 34.19\)). The difference score for mean fixation durations was calculated accordingly (duration difference: \(M = 27.97, SD = 63.56\)). These two values were entered as predictors in a linear regression model with performance in test 2 as outcome. As expected, the higher the increase in relative attention on the picture across cycles (higher count difference), the better was the performance in test 2, \(\beta = .46, p = .001\).

### Table 3

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Multimedia (Cycle 1)</th>
<th>Multimedia (Cycle 2)</th>
<th>Multimedia (Combined)</th>
<th>Text only (Cycle 1)</th>
<th>Text only (Cycle 2)</th>
<th>Text only (Combined)</th>
</tr>
</thead>
<tbody>
<tr>
<td>JOLs (min. – 0, max. 100)</td>
<td>73.48 (14.84)</td>
<td>80.44 (16.71)</td>
<td>76.96 (15.31)</td>
<td>58.81 (18.58)</td>
<td>67.04 (20.53)</td>
<td>62.93 (18.14)</td>
</tr>
<tr>
<td>Calibration scores</td>
<td>44.79 (11.77)</td>
<td>44.10 (14.47)</td>
<td>44.44 (12.05)</td>
<td>35.12 (17.78)</td>
<td>34.28 (18.02)</td>
<td>34.70 (15.42)</td>
</tr>
<tr>
<td>Learning outcomes</td>
<td>36.64 (5.23)</td>
<td>21.08 (5.98)</td>
<td>18.96 (4.28)</td>
<td>13.74 (5.87)</td>
<td>19.00 (4.30)</td>
<td>16.37 (4.68)</td>
</tr>
<tr>
<td>Pictorial matching scores</td>
<td>17.70 (1.75)</td>
<td>17.70 (1.75)</td>
<td></td>
<td>14.28 (2.88)</td>
<td>14.28 (2.88)</td>
<td></td>
</tr>
<tr>
<td>Study times (in seconds)</td>
<td>334.05 (91.26)</td>
<td>254.46 (91.37)</td>
<td>294.26 (77.22)</td>
<td>311.32 (99.11)</td>
<td>261.71 (65.21)</td>
<td>286.52 (67.88)</td>
</tr>
</tbody>
</table>
tion duration on the text-AOI as dependent variable revealed no action (see Fig. 4). Similarly, a 2 and condition (multimedia vs. text only) as factors, and mean effort investment into text reading. The ANOVA with cycle (1 vs. 2) revealed main effects of cycle, cycle (1 vs. 2) and condition (multimedia vs. text only) as factors. Concerning cycle, there was a main effect of cycle (1 vs. 2, $F(1,44) = 17.72$, $p < .001$, $\eta^2_g = .28$, and no interaction, $F < 1$. Hence, mean fixation durations remained the same across cycles, but they were generally longer in the text-only than in the multimedia condition. Neither total fixation count nor mean fixation durations on text in cycle 2 were related to performance in test 2 in the text-only condition, both $\beta_1 < .10$, both $t < 1$.

With respect to the total amount of attention spent on text and pictures in both conditions (measured via total fixation times on text- and picture-AOI), a $2 \times 2$ repeated-measures ANOVA with cycle (1 vs. 2) and condition (multimedia vs. text only) as factors revealed main effects of cycle, $F(1,44) = 16.55$, $p < .001$, $\eta^2_g = .27$, and condition, $F(1,44) = 4.18$, $p = .047$, $\eta^2_g = .09$, and no significant interaction, $F(1,44) = 3.29$, $p > .05$. Hence, aside from a decrease across cycles, students in the multimedia condition had longer overall fixation times than students in the text-only condition. Longer overall fixation times were not correlated to better learning outcomes in the multimedia, $r = -.26$, $p > .05$, or the text-only condition, $r = -.07$, $p > .05$.

### 3.3. Discussion

As in Experiment 1, there was support for the multimedia effect (Butcher, 2014), for the multimedia heuristic (Serra & Dunlosky, 2010), and for better test scores and faster learning in the second than in the first study-test cycle.

As in Experiment 1, lower performance in test 1 was associated with longer study times in cycle 2 compared to cycle 1. This relationship could not be explained by changes in confidence levels, hence not confirming Hypothesis 6, but supporting Hypothesis 1 of the ‘adaptation to task demands’ account. Supporting Hypotheses 2a and 2b of the ‘adaptation to task demands’ account, fixation count on text decreased more strongly than fixation count on the picture and mean fixation durations on the picture increased across cycles in the multimedia condition (see Fig. 4). Partially supporting Hypothesis 2c, the relatively higher fixation count on the picture than on the text (but not the longer mean fixation durations) in cycle 2 was related to better study outcomes, suggesting that reducing the bias of over-reliance on text in multimedia learning was associated with better learning success. Supporting Hypothesis 3a, students in the text-only condition attended longer to text than in the multimedia condition only in study-test cycle 2 (see Fig. 4). Partially supporting Hypothesis 3b, students in the text-only condition had longer mean fixation durations than students in the multimedia condition, suggesting that they invested more effort reading the text. However, mean fixation durations were constantly higher in the text-only than in the multimedia condition (Eitel, Scheiter, & Schüller, 2013), and not especially in study-test cycle 2. Moreover, these measures were unrelated to learning outcomes, hence not supporting Hypothesis 3c.

As in Experiment 1, stable overconfidence across cycles was found that was even more pronounced with multimedia, hence supporting Hypothesis 4a but not 4b of the ‘underconfidence-with-practice’ account (see Fig. 3, right panels). Moreover, as in Experiment 1, and not supporting Hypotheses 5a and 5b, the multimedia effect remained unchanged across cycles, and study times were not longer in the multimedia than in the text-only condition in cycle 2.

### 4. General discussion

Two experiments were conducted to investigate the multimedia effect and how it is affected by repeated studying and testing. Students learned either with text only or with text and pictures (multimedia), were tested on the learning contents and repeated this study-test cycle. On an outcome-level, results revealed a robust multimedia effect across study-test cycles. On a process-level, two hypotheses were derived as to how two types of biases in multimedia learning — over-reliance on text processing and over-confidence — could be reduced when studying and testing are repeated.

First, because after the initial study-test cycle, students know what to expect from the next cycle, they should adapt their learning to the demands of the initial test (‘adaptation to task demands’). Hence, they should process text and pictures more deliberately to extract specific information needed for the test, thereby reducing the bias of over-reliance on text processing in multimedia learning. Second, overconfidence due to multimedia should be reduced when studying and testing are repeated, because success in retrieving information in the first test serves as a valid cue to judge performance in test 2 (‘underconfidence-with-practice’). This should trigger more effortful processing of multimedia materials in cycle 2, in turn fostering learning outcomes.

---

2 There was one outlier with respect to fixation count on text in the text-only condition, who was excluded from the analysis. Including this outlier in the analysis would have led to a violation of the ANOVA assumption of normally distributed data.
4.1. Evidence for the ‘Adaptation to task demands’ account

Results from both experiments support the hypothesis of adaptation to task demands. Students showed in both conditions (multimedia and text-only) that they adapted their learning in cycle 2 to the demands of test 1, because the more difficulties they had solving test 1 (lower performance), the more time they allocated for studying in cycle 2 (compared to cycle 1). Hence, they regulated their learning in response to task demands (Pieschl et al., 2012). This regulation was effective, as shown by robust improvements in learning outcomes from cycle 1 to cycle 2 in the present experiments. Learning outcomes improved to a similar degree in the multimedia and in the text-only condition across cycles, but the processes that led to these improvements were quite different, as shown by the eye movement analysis of Experiment 2. This analysis is crucial as it sheds more light on the processes underlying the multimedia effect and how it develops across study-test cycles.

The eye movement analysis first revealed that students’ number of fixations devoted to text processing decreased more steeply in the multimedia than in the text-only condition from cycle 1 to cycle 2. In the text-only condition, students’ attention on text hardly decreased across cycles (see Fig. 4). As a higher number of fixations should reflect a higher degree of cognitive processing (adaptation & Just. 1980; Rayner et al., 2006), one might conclude that students adapted their processing to task demands in cycle 2 by spending relatively more time carefully re-reading text than students in the multimedia condition, for instance by mentally imagining the contents.

Second, in line with prior research (Eitel, Scheiter, & Schüler, 2013; Folker et al., 2005; Lenzner et al., 2013), students in the multimedia condition attended longer to the text than to the pictures in the first study phase, suggesting that compared to text pictures were hardly processed. One might conclude that because students did not know the exact demands of the upcoming test, they relied on their general text-based learning strategies (Lorch et al., 1993), resulting in the bias of over-reliance on text processing in multimedia learning. Over-reliance on text processing was reduced in cycle 2, presumably because knowing the demands of the first test enabled students to process text and pictures more deliberately to extract the information needed in the upcoming test. This was reflected in relatively more fixations on the picture and in longer mean fixation durations on the picture. Relatively more fixations on the picture were in turn related with better performance in test 2, which is in line with prior research (Eitel, Scheiter, & Schüler, 2013; Schonke et al., 2009). Moreover, in line with prior research, the present results suggest that interventions to support a more extensive picture processing, in turn, also support better learning outcomes (Scheiter & Eitel, 2015), indicating that the potential of pictures to foster learning is better exploited the more extensively they are processed. Therefore, repeating the study-test cycle can serve as an intervention to optimize processing of multimedia instructional materials by reducing the bias of over-reliance on text processing.

However, even though mean fixation durations on the picture increased across cycles, this increase was not related to better performance, suggesting that the increased effort into processing of the picture did not pay off at an outcome-level. This missing link may be explained by referring to two different processes that underlie longer mean fixation durations. First, higher mean fixation durations can result from more elaboration and memorization processes (Henderson & Hollingworth, 1999), which would be reflected in a positive relationship between mean fixation durations and performance. Second, higher mean fixation durations can result from problems in deciphering, identifying, or understanding the picture with its elements, and therefore would be negatively related to performance.

Taken together both processes, the correlation between mean fixation durations and performance would be zero, as was observed in the present research. One promising approach to empirically support this claim is to later ask students about their processing behavior by exposing them to their eye movements recorded during the learning episode (cued retrospective reporting; Van Gog, Paas, Van Merriënboer, & Witte, 2005). This could be subject to further research.

Another interesting direction for further research would be to determine the boundary conditions under which beneficial effects of adaptation to task demands appear. Specifically, in addition to knowing what was tested, after completing the first test students in the present experiments also knew how it was tested (i.e., in an open question format). Thus, the present results are in line with the test expectancy literature (Lundeberg & Fox, 1991), suggesting that knowing the format of a test fosters performance (open questions vs. verification). Moreover, in addition to knowing the format of a test, after completing the first of two identical tests students knew what knowledge would be asked in the second test (memory vs. inference; Thiede et al., 2011). These can be considered two different influences on subsequent studying and testing, and entangling the unique contributions to better performance in the second test should be subject to further research in the context of multimedia learning. To this end, further research may compare the effects of using two identical tests in both study-test cycles (as in present research) to using tests in different formats and/or tests that assess different types of knowledge in the first and second study-test cycle. According to the ‘adaptation to task demands’ hypothesis, repeated studying and testing should be the more helpful the more similar the tests in both cycles. Hence, performance should increase more steeply with identical tests than with tests that differ in their format and/or in the type of assessed knowledge.

4.2. Reasons for overconfidence despite practice

In line with prior research (Serra & Dunlosky, 2010), the present experiments revealed higher confidence through multimedia in the first study-test cycle. But unlike expected from the ‘underconfidence-with-practice’ account, practicing retrieval in test 1 did not lead to a reduction in over-confidence due to multimedia. Rather, confidence levels remained as inflated as in cycle 1, and tended to get even more inflated in cycle 2 in one of the present experiments, suggesting that the bias of overconfidence in multimedia learning was not reduced. Hence, underconfidence-with-practice effects, as obtained in more basic memory research using word lists or short texts (Carpenter & Olson, 2012; Kornat et al., 2002), were not found to be applicable to a more complex instruction consisting of expository text (with pictures). One reason for this discrepancy was the subtlety of feedback in the present studies.

To comply with previous research in multimedia learning (Mayer, 2014), open questions without direct feedback were used to assess knowledge in the present research. This might have prevented a confidence reduction, because students might not have been aware of how well they performed on the first test, and thus, of their potential knowledge gaps and misunderstandings. Thus, their subjective success in answering the test questions deviated from the objective success, so that their subjective success in answering test 1 was not a valid cue on which to base subsequent study regulation. Since inaccurate monitoring leads to inadequate regulation (metamemory framework; Nelson & Narens, 1990), students remained (over-)confident and did not invest more study effort in cycle 2 than in cycle 1, as shown by higher JOLs, shorter
study times, and robust mean fixation durations across cycles. In addition, students might have used their JOL in the first cycle as an anchor for their JOL in the second cycle (Zhao & Linderholm, 2008), as indicated by high correlations between JOLs in both cycles and experiments (.81 ≤ r ≤ .88, both p<.001). And since the same materials were studied repeatedly, students were likely to become more familiar with the materials in cycle 2, potentially leading to JOLs in cycle 2 that were higher than the anchor-JOLs from cycle 1.

Hence, the absence of direct feedback might have been responsible for overconfidence despite retrieval practice; providing feedback might result in subsequent confidence reduction that might, in turn, result in higher study effort. Feedback could be provided either by telling students how many of their answers were correct (performance feedback) or by presenting them with the sample solutions to the test (content feedback). Although both types of feedback are likely to have substantial effects on subsequent learning, one has to be aware of the methodological difficulties associated with this procedure. Specifically, if students are told how many answers were correct, then they might take the concrete number provided by the feedback as a basis for their subsequent confidence judgment so that a decrease in JOL magnitudes from cycle 1 to cycle 2 might not reflect real confidence reduction but rather the switching of the anchor (from JOL1 to feedback 1). If sample solutions to the questions are provided in test 1, then test 2 would need to be very different from test 1, because otherwise students solely would have to recall the sample solutions from test 1 to perform well in test 2. Constructing a test 2 that is different from test 1, but at the same time identical in its difficulty level would pose a challenge. In conclusion, investigating the effects of feedback on confidence (reduction), and potentially more accurate calibration in multimedia learning is interesting but also challenging on a methodological level.

4.3. Limitations and conclusions

Aside from many consistent results between the two present experiments, results for study times and calibration scores are partially inconsistent between experiments. Specifically, whereas study times decreased more strongly from cycle 1 to cycle 2 in the multimedia than in the text-only condition in Experiment 1, study times decreased similarly in both conditions in Experiment 2. However, whereas study times did not differ significantly in how they decreased between conditions in Experiment 2, the pattern of study time differences is consistent with Experiment 1 on a descriptive level; that is, study times decreased more strongly in the multimedia than in the text-only condition. The difference did just not reach significance in Experiment 2. Results for calibration scores differ somewhat between experiments in that they increased across cycles in Experiment 1 but remained unchanged across cycles in Experiment 2. Interestingly, results for calibration scores (JOLs—learning outcomes) differ between experiments even though results for JOLs are consistent; that is, JOLs were consistently higher due to multimedia, and increased across cycles in both experiments. Hence, inconsistencies in results for calibration can be explained by the sole fact that learning outcomes improved even more strongly in Experiment 2 than in Experiment 1 across cycles (see Fig. 3, left panels), while JOLs increased to a similar extent in both experiments. This suggests that the same processes were triggered by repeated studying and testing in both experiments — students learned better and perceived their learning as being better in cycle 2 than in cycle 1.

In the present research, hypotheses addressed how multimedia learning develops across cycles on several different levels; that is, on the levels of metacognitive awareness (JOLs), cognitive processes (eye movements, study times), learning outcomes, and how they are related to each other (calibration; performance predicted by attention on picture). Hence, hypotheses addressed learning outcomes and their relations to (meta-)cognitive processes in general; that is, predictions were made only for the composite scores for learning outcomes, and it was not distinguished between recall and transfer (troubleshooting of the system). This would be an interesting issue for further research that takes fewer levels into account and focuses on this aspect. For further research in this area the use of eye tracking is recommended, because it can provide valuable insights into students’ processing behavior in learning situations on a fine-grained level.

To warrant a high internal validity, the same frequently used instructional materials about the structure and functions of a toilet flushing system were used in both of the present experiments (Mayer et al., 2005; Paik & Schraw, 2013). Hence, I cannot exclude idiosyncratic features of the materials to having influenced the effects obtained within the present research, and call for further research using different types of instructional materials to allow for conclusions regarding their generalizability. Moreover, further research could also investigate potential overconfidence-effects when learning with text and dynamic pictures (animations, video) because especially moving pictures are assumed to inflate metacomprehension judgments (illusion of understanding; Kühn et al., 2011; Paik & Schraw, 2013). Nonetheless, using materials from the domain of mechanics, the present research replicated overconfidence-effects found with materials from the domain of meteorology (Serra & Dunlosky, 2010), indicating the potential generalizability of the results.

To conclude, the present research revealed that a basic effect from memory research could be applied to learning with more complex materials in the multimedia context; that is, repeated studying and testing fostered performance. Moreover, performance was fostered to a similar degree when learning with both text-only and multimedia, showing that the multimedia effect was robust across cycles and suggesting that effects of the instructional design obtained in the first cycle of learning and testing extend to subsequent cycles. In addition, the present research suggests that testing can lead to desirable effects on a processing level when relearning with multimedia; that is, to a relatively higher attention on the pictures.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.learninstruc.2015.10.003.

References


