June 2013

Ackerman, R., & Leiser, D. (in press). The effect of concrete supplements on metacognitive regulation during learning and open-book test taking. *British Journal of Educational Psychology.*

The Effect of Concrete Supplements on Metacognitive Regulation during Learning and Open-book Test Taking

Rakefet Ackerman[1]* and David Leiser[2]

[1] Technion–Israel Institute of Technology, Haifa, Israel
[2] Ben-Gurion University of the Negev, Beer-Sheba, Israel

*Requests for reprints should be addressed to
Rakefet Ackerman
Faculty of Industrial Engineering and Management
Technion–Israel Institute of Technology
Technion City, Haifa, Israel, 32000
ackerman@ie.technion.ac.il

Abstract

Background. Previous studies have suggested that, when reading texts, lower achievers are more sensitive than their stronger counterparts to surface-level cues, such as graphic illustrations, and that even when uninformative, such concrete supplements tend to raise the text's subjective comprehensibility.

Aims. We examined how being led astray by uninformative concrete supplements in expository texts affects achievement. We focused on the mediating role of metacognitive processes by partialling out the role of cognitive ability, as indicated by SAT scores, in accounting for the found differences between higher and lower achievers.

Sample and method. Undergraduate students studied expository texts in their base versions or in concrete versions, including uninformative supplements, in a within-participant design. The procedure had three phases: studying, open-book test taking, and reanswering questions of one's choice.

Results. Overall, judgments of comprehension (JCOMP) were higher after participants studied the concrete than the base versions, and the participants benefited from the open-book test and the reanswering opportunity. An in-depth examination of time investment, JCOMP, confidence in test answers, choice of questions to reanswer, and test scores indicated that those whose metacognitive processes were more effective and goal-driven achieved higher scores.

Conclusions. The effectiveness of metacognitive processes during learning and test taking constitutes an important factor differentiating between higher and lower achievers when studying texts that include potentially misleading cues.

People often study textual materials in their educational and professional lives in a self-regulated manner. Self-regulated learning (SRL) is an active, constructive process whereby students regulate the effort they invest in learning so as to achieve self-set learning goals (Azevedo, 2009; Bjork, Dunlosky, & Kornell, 2013; Efklides, 2011; Pintrich, 2000). While reading a textbook chapter, for example, students monitor how well they understand the material and decide whether to invest more study time, study some of the material anew, apply specific learning strategies, or stop studying, according to their motivation for success (Greene & Azevedo, 2007; Hacker, 1998; Pressley, 2002; Winne & Hadwin, 1998; Zimmerman, 2008). We refer to this combination of (a) setting a desired knowledge level, (b) monitoring, and (c) control over time investment directed toward reaching this goal as *Metacognitive Learning Regulation* (MLR; see Ackerman & Goldsmith, 2011; Ackerman & Lauterman, 2012). This definition is meant to differentiate these aspects of learning regulation from others aspects often discussed under the umbrella of SRL, such as metacognitive knowledge, reflection about previous task experiences, and choice among learning strategies (see Brown, 1987; Zohar & Dori, 2012).

Previous studies have confirmed that effective MLR depends on the reliability of the learners' ongoing subjective monitoring of their acquired knowledge level (Metcalfe & Finn, 2008; Thiede, Anderson, & Therriault, 2003). Disturbingly, however, studies have also found that people studying from texts are not very good at assessing their actual level of knowledge and predicting their performance at test (see Dunlosky & Lipko, 2007, and Maki, 1998, for reviews). Studies which examined individual differences have pointed out that stronger students monitor their knowledge more reliably than their weaker counterparts (e.g., Bol & Hacker, 2001; Maki &

Berry, 1984). One explanation for this monitoring difference is that in the course of their learning, weaker students are more likely to base their monitoring of comprehension on surface-related cues (Thiede, Griffin, Wiley, & Anderson, 2010). Similarly, poor problem solvers were found to be more misled by superficial aspects of the problems (Novick & Sherman, 2008). These findings may lead to an association between general cognitive ability and susceptibility to misleading heuristic cues. However, recently Ackerman, Leiser, and Shpigelman (2013) found that when the task is particularly challenging, even students with top SAT scores are misled by superficial characteristics of the study materials. This suggests that general ability of this type is not predictive of susceptibility to such misleading cues. In the present study, we aimed to establish empirically that, in regular text learning tasks, individuals more prone to the misleading effects of surface-level cues on their metacognitive monitoring and regulation of study time would achieve lower scores in comprehension tests, regardless of their general ability as reflected by SAT scores.

One type of surface-level cue that has been found to underlie metacognitive monitoring for text learning is the inclusion of supplements, such as illustrations. Well-designed illustrations clearly enrich understanding (e.g., Ainsworth, 2006; Carney & Levin, 2002; Mayer, 2003). This is often done by displaying relations between several elements, whether causal, spatial or temporal (e.g., Mayer, 2003; Serra & Dunlosky, 2010). In the presence of such elaborative illustrations, assessing higher knowledge is well justified. However, evidence is accumulating that including illustrations in texts does not always improve, and under certain conditions may even impair, performance at test (Harp & Mayer, 1998; Prangsma, Boxtel, Kanselaar, & Kirschner, 2009).

One explanation for this performance impairment is that redundant graphical and textual contents in texts increase the cognitive load on learners (Chandler & Sweller, 1991; Mayer, Heiser, & Lonn, 2001). Two recent studies have examined metacognitive bias as a complementary explanation. Serra and Dunlosky (2010; Experiment 2) tested participants on an expository text explaining how lightning storms develop. When a photograph of lightning that offered no elaborative value was included, predictions of performance rose markedly, but study time and actual test scores were not significantly different. Looking beyond pictorial illustrations, previous findings have suggested that concrete textual content is associated with subjective comprehensibility (Sadoski, Goetz, & Fritz, 1993). While these studies did not examine the implications on subsequent MLR decisions (e.g., whether to restudy the text), overestimation of knowledge can be expected to misdirect such regulatory decisions (Metcalfe & Finn, 2008). Ackerman et al. (2013) examined metacognitive effects of incorporating non-informative illustrations of relevant objects in explanations of solutions for particularly challenging logic problems. They too found that the presence of uninformative illustrations enhanced participants' metacognitive judgments and did not affect study time. In their study, the misleading effect was even stronger: the illustrations reduced success rates in solving near-transfer problems.

Together, these studies suggest that when concrete supplements, either graphical or textual, do not contribute to actual knowledge acquisition, their concreteness may inflate the learner's metacognitive judgment such that the learner assesses his/her knowledge level as higher than will ultimately be justified by performance. Such inflated judgments are expected to hinder recognition of the need to invest additional cognitive effort, as suggested by the cognitive load explanation (Chandler & Sweller,

1991; Mayer, Heiser, & Lonn, 2001). Following this analysis, we hypothesized that a comparison using concrete and base versions of expository texts would enable us to empirically establish an association between the extent to which individuals are prone to the misleading effects of concreteness and their achievement level.

We used two versions of each expository text in our study. The concrete versions included the same wording as the base versions, supplemented by illustrations presenting isolated objects mentioned in the text, with no elaborative value (e.g., spinach leaves), and a few brief verbal examples and concrete associations (e.g., mentioning Popeye the Sailor Man in the context of spinach; see Appendix). The supplements were intended to increase the text's concreteness without contributing to comprehension, and this was verified by a manipulation check (see Materials section). The comprehensive set of MLR measures included objective variables (time and test scores), subjective variables (monitoring of comprehension and confidence), and indices derived from the relationships among them, as detailed below. Importantly, we took care to focus on individuals' differing propensity to be misled by concreteness as the underlying source of the performance difference among participants by partialling out the performance variability explained by SAT scores.

People have various regulatory choices during learning, including the option of returning to portions of the material which they assess as less well understood (Thiede et al., 2003). To extend our MLR measures beyond those previously analyzed, we used an inference test and employed an open-book testing procedure. While rereading by itself benefits learning outcomes, open-book testing was found to improve learning even more (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008). Previous studies have suggested that open-book testing encourages higher level thinking than

closed-book testing (Feller, 1994; see Heijne-Penninga, Kuks, Hofman, & Cohen-Schotanus, 2011, for a review), and this focus of learning presumably also directs metacognitive monitoring towards assessing high-order understanding (Thiede, Wiley, & Griffin, 2011). The present study is the first to integrate metacognitive analysis of the study and test taking phases in the presence of misleading surface-level cues.

One underlying hypothesis guiding us was that concrete supplements increase judgments of comprehension (JCOMP; cf. Zaromb et al., 2010) at the end of the study phase, as previously found with both JCOMPs and predictions of performance (Ackerman et al., 2013; Serra & Dunlosky, 2010). However, what should we expect in relation to the metacognitive regulation of time, which is another central aspect of MLR? Regarding the study phase, one might predict that as people perceive concrete texts to be more comprehensible, they will rate their understanding as higher and therefore invest *less* study time, resulting in a negative correlation between time and judgment. This is expected to raise metacognitive judgments but to yield lower performance than could have been achieved by regulating time according to well-calibrated judgments (Nelson & Narens, 1990; see Ackerman & Goldsmith, 2011, for an illustration). However, a positive correlation between time and judgment also makes sense. Koriat, Ma'ayan, and Nussinson (2006; see also Koriat, Ackerman, Adiv, Lockl, & Schneider, in press) assigned lower and higher incentives to paired associates presented for memorization. The items assigned higher points were studied longer, and participants expected to succeed better in remembering them. The additional time invested indeed proved effective. Thus, elevated judgments have been empirically associated with both a decrease and an increase in allocated study time.

What are we to make of the observation that metacognitive judgments sometimes appear to be negatively correlated with study time, and sometimes positively correlated with it? Koriat et al. (2006) suggested an explanation for this puzzle, by differentiating between *data-driven* and *goal-driven* effort investment. In their view, negative correlation characterizes study sets which introduce variability in the items' perceived difficulty. Data-driven effort investment leads to a situation where difficult materials take longer to study, yet carry lower chances of success than easier materials. Goal-driven effort investment characterizes cases in which the motivation for success is stronger for some items than for others, as with the above-mentioned higher and lower point values associated with different items. Motivation to succeed leads the learner to invest more time in an attempt to improve performance than when motivation is lower (Fisher & Ford, 1998; Grant & Dweck, 2003; Pintrich, 2000). This effort investment results in a positive correlation between invested time and metacognitive judgments.

Considering the findings of equivalent time investment in learning tasks with and without illustrations despite differences in metacognitive judgments (Ackerman et al., 2013; Serra & Dunlosky, 2010) in light of Koriat et al.'s (2006) theorizing raises the possibility that these results reflect the use of different processes by different participants. In other words, although metacognitive judgments were higher overall for the illustrated texts, it may be that some participants invested more time in the illustrated versions and others invested less, with the two groups basing their metacognitive monitoring on goal-driven and data-driven effort, respectively. Following this analysis, our hypothesis in the current study was that JCOMPs would be higher in the presence of non-informative concrete supplements across the entire

sample. However, lower achievers would be misled to consider concrete texts as easier and, therefore, to engage in data-driven effort investment, which would lead them to terminate study earlier than for base texts, while higher achievers would recognize the need for extra cognitive effort, and so improve their scores.

A further question of interest is whether being misled by the concrete supplements would affect regulation of test taking. In particular, we explored whether shorter study time in the presence of the concrete supplements would be compensated for by the opportunity afforded by the open-book testing procedure (see Year 2 group of Heijne-Penninga et al., 2011). Moreover, we explored the effects of processing during the test phase on the test-related MLR measures detailed below.

The experiment

Undergraduate students studied two texts each, one in its concrete version and one in its base version. Immediately afterward, they answered open-ended questions with open books. Finally, they were allowed to choose two questions to reanswer. The use of within-participant manipulation of the text versions allowed a focus on the sensitivity of the various MLR measures to the presence of concrete supplements at the individual level. Moreover, to control for other relevant factors that affect achievements at test, such as general reading ability and working memory (Griffin, Wiley, & Thiede, 2008), we also collected SAT scores. This allowed us to focus on the role of MLR sensitivity to the surface-level manipulation of concreteness in explaining differences among participants in their ultimate scores.

MLR was examined using the following set of dependent variables. For the study phase we measured *self-paced study time* and *JCOMP*. During the test phase we measured overall *test taking time* and for each answer, *confidence* and *score*. Two

measures were used to indicate the accuracy of the confidence ratings. Absolute monitoring accuracy, or *calibration,* is the difference between a learner's mean metacognitive judgment and test score and reflects under- or overconfidence (e.g., Ackerman & Goldsmith, 2011; Hacker, Bol, & Bahbahani, 2008). Relative monitoring accuracy, or *resolution,* measures the extent to which metacognitive judgments distinguish correctly between better and more poorly performed tasks; it is calculated using a within-participant Goodman-Kruskal gamma correlation (see Nelson, 1984; e.g., Thiede et al., 2003). Previous studies have suggested that increasing the depth of processing during the learning phase through manipulations such as asking participants to write summaries, generate keywords, or fill in missing letters enhances the resolution of comprehension judgments (Anderson & Thiede, 2008; Maki, Foley, Kajer, Thompson, & Willert, 1990; Thiede, Dunlosky, Griffin, & Wiley, 2005). We are not aware of similar findings regarding confidence ratings, but in the present study, the open-book test format allowed participants in-depth processing of the texts when answering each question, and this was expected to enhance the resolution of confidence ratings in a similar manner to that found for the learning phase.

The reanswering phase contributed additional information as to the participants' MLR. The strength of the relationship between a confidence judgment and a decision to act in keeping with that judgment (in our case, the decision to reanswer a question) is called *control sensitivity* (Koriat & Goldsmith, 1996). Control sensitivity is important for practical decisions, such as when a choice of questions is allowed in tests, or when learners must decide whether to seek help. In memory-related tasks this measure is usually at its ceiling for young and healthy adults and

weaker in older adults and people with mental illnesses (Koren, Seidman, Poyurovsky, Goldsmith, Viksman, Zichel, & Klein, 2004; Pansky, Koriat, Goldsmith, & Pearlman-Avnion, 2009). However, recent findings suggest that for text learning there may be variability in control sensitivity even among young and healthy participants (Ackerman & Goldsmith, 2011). Here, the within-participant gamma correlation between each confidence rating and the decision whether to reanswer the question indexed control sensitivity for that test.

Method

Participants

Forty-four participants (mean age 23.4) participated in the experiment for course credit. They were recruited from the social science departments of one university and two colleges. The participant recruitment notice specified that participants should not have any type of learning disability, and students who reported having learning disabilities on their personal data form were excluded from participating. All participants were either native Hebrew speakers or had studied in Israel in Hebrew at least since the first grade. Their self-reported Israeli SAT scores ranged between 444 and 768 (M = 628, SD = 86). Scores on this test range from 200 to 800, the national mean is 540, 15% nationally score above 650, and 5% score above 700. An additional 85 participants drawn from the same population provided the samples for Pretests A and B (described below).

Materials

The materials comprised six expository texts (300–600 words long). Six texts were chosen to ensure that we would examine general effects, independent of particulars of a specific text. The texts were printed on one or two pages, and dealt with topics in

which participants were not expected to be highly knowledgeable: urban air pollution, the process of balding, the planet Mars, the nutritional value of spinach, ancient Petra in Jordan, and the wild boar's adaptation to changes in its environment. For each text we prepared two versions: concrete and base. The concrete versions included the same wording as the base versions, together with colorful photographs or drawings of isolated objects mentioned in the texts, and two to four brief textual examples or common concrete associations (see Appendix). Importantly, the supplements offered no elaborative value for understanding the text's main ideas. The concrete versions included about 10% more words than the base versions.

Pretest A was a manipulation check that examined the impression generated by the text. Each participant ($N = 48$) received two texts, one in its base version and the other in its concrete version, with their order counterbalanced across participants. The questionnaire comprised eighteen questions to be answered on a 5-point scale ("not at all" to "very much"), of which thirteen were filler questions (involving frequency of words, length of lines, characteristics of the topic, etc.) and three were target questions: "How concrete is the text?", "To what extent did the author use concrete examples and illustrations in this text?", and "To what extent do you think the author could have presented the text in a more concrete way?" (reverse-scored). A paired comparison between the mean concreteness-related ratings indicated that the versions with the supplements were indeed experienced as more concrete ($M = 3.5$, $SD = 0.7$) than the base versions ($M = 3.1$, $SD = 1.0$), $t(47) = 2.74$, $p < .01$. The remaining two questions controlled for perceived interest and difficulty: "How interesting/difficult is the text for a student who needs to study it for a comprehension

test?" (M = 3.4, SD = 1.1; M = 2.4, SD = 1.2, for interest and difficulty, respectively). Both control questions yielded no difference between the versions, both ts < 1.

Pretest B confirmed that the texts led to new knowledge acquisition with no difference between the versions, and provided the basis for selecting questions for the main study. For each text we prepared ten open-ended questions, all requiring inference from information included in more than one sentence. The participants (N = 37) first answered the ten questions based on their general knowledge, and then again with the text in hand. To ensure reliability in the scoring, two judges rated the answers using a predefined detailed scoring norm which specified the points to be given for each element of the correct answer (each answer could earn up to ten points). For each test, the final score for each question was the average of the judges' ratings (r = .81, p < .0001), expressed as a percentage. Lack of preliminary proficiency was verified (M_{score} = 29%; SD = 12.1), and the clarity of the texts and questions was confirmed, as with the text in hand the participants could answer the questions reasonably well (M_{score} = 74%; SD = 10.2). Importantly, there was no ceiling effect, allowing us to examine effects on test scores. Questions were selected for the main study according to the following criteria: first, we eliminated questions for which participants' mean score with the text in hand was below 20%, indicating that the question or the solicited information was unclear. We further eliminated questions for which the mean score was above 95%, as being too easy for our purposes. Among the remaining questions, we chose those for which new knowledge acquisition was the largest (i.e., those with the largest difference between scores on the preliminary and subsequent tests; for all questions chosen the difference was at least 30%). For all six texts at least five questions passed the screening procedure. No performance differences were

found between the concrete and base versions, either when including all ten questions or when including only those questions selected for the main study (both *t*s < 1).

Procedure

Experimental sessions were conducted in regular classrooms for five to twelve participants at a time. Each participant studied two texts randomly chosen out of the pool of six, one in the concrete version and the other in the base version, with their order counterbalanced across participants. The instructions were as follows: "You will first be asked to read the entire article in depth, so as to gain a good understanding of its contents. Later on, you will be asked 5-7 open-ended inference questions, most of which relate to information dispersed throughout the text rather than presented in a single location. The questions require a thorough understanding of the article, so please read the article carefully with this in mind. You are welcome to mark up the text and take notes freely on the paper. There is no need to memorize details, as the text will be in front of you and you will be able to consult it when you answer the questions. Your task is to understand the ideas expressed in the text and learn its contents in one study cycle. There is no time limit—study until you feel that you know the materials well." Thus informed, participants could in principle opt to engage in shallow processing during the study time and count on getting back to the texts during the test. Since it was not possible to prevent this strategy altogether, our analyses examined whether this was the case.

JCOMP ratings were collected as soon as participants were ready for the test on a given text. Participants marked their judgments by drawing a vertical line on a 0–100% horizontal scale in response to the question "To what extent do you feel that you understand the text?" They then turned to the open-book test. While taking the

test, participants indicated their confidence in each answer (0–100%). Finally, after completing the test, they were allowed to reanswer up to two questions of their choice by restudying the text. Participants wrote their revised answers on a separate sheet of paper. The time (HH:MM) was documented by the participants at each phase of the experiment.

Results

Below, we first present the overall test scores and the way we used participants' achievement for analyzing the susceptibility of MLR to the presence of concrete supplements. We then analyze the effects of the text versions on MLR components.

The effect sizes of the t-tests were measured by Cohen's *d,* following commonly accepted guidelines that refer to effect sizes of 0.20, 0.50, and 0.80 as small, medium, and large, respectively (based on Cohen, 1988; see Fritz, Morris, & Richler, 2012). Effect sizes of ANOVAs were measured by η_p^2 (partial eta-squared) with .01, .06, and .14 representing small, medium, and large, respectively (based on Cohen, 1988; see Richardson, 2011).

Test scores

Two judges, blind to the text version, graded the answers based on the detailed scoring norms prepared in advance of Pretest B. After verifying interjudge reliability ($r = .85$, $p < .0001$), we took the mean score from the two judges as the test score for each answer (0-100%), and averaged these to produce an overall initial score for each participant for each text. Scores were then calculated for the final set of answers (after participants had the chance to reanswer two questions). Across all texts and participants, the mean initial score ($M = 53.1\%$, $SD = 13.3$) was lower than the final score ($M = 58.5\%$, $SD = 13.5$), $t(43) = 4.83$, $p < .0001$, $d = 0.73$. This means that

participants effectively used the opportunity to reanswer two questions, and improved their scores by about five points on average. This finding is important, as it highlights the fact that the participants did not judge their work during the test to be exhaustive, and made an effort to improve their achievement (see further support for this conclusion under Time regulation below).

The mean final test scores across the two versions were markedly bimodal, suggesting that the participants were naturally divided into two groups. Accordingly, to examine the effect of concreteness and whether it differed for different levels of performance, we used the median of participants' mean final test scores to categorize participants as either lower (N = 22, M = 47.3%, SD = 6.9) or higher achievers (N = 22, M = 69.8%, SD = 7.8) (for a similar approach see Bol & Hacker, 2001; Maki & Berry, 1984). There was a normal distribution of scores within each half over both versions and for each version separately (all ps > .70 by Kolmogorov-Smirnov Z test). Table 1 presents the means (and SDs) of the measures and the results of the statistical comparisons between the groups and the versions.

Having thus classified the participants, we investigated how the text version affected the test scores at each achievement level. To this end, we ran a two-way Analysis of Variance (ANOVA) examining the effect of Achievement (low vs. high) × Text version (concrete vs. base) on the final test scores. This yielded, of course, a large main effect of achievement. Further, there was also a marginal effect of the text version with a medium effect size and, notably, a significant interactive effect (see Table 1): The higher achievers scored higher for the concrete than for the base versions. The lower achievers, in contrast, achieved similar scores for the two versions. A similar interactive effect was found for the initial scores, before the

reanswering phase (see Table 1). The following analyses attempt to expose the MLR processes associated with these differing effects of the text versions on the two groups of participants.

Table 1. Means and SDs and ANOVA results regarding the effects of achievement level and text version on the various Metacognitive Learning Regulation (MLR) measures.

	Higher achievers				Lower achievers				$F(1, 42)$[1]								
	Concrete		Base		Concrete		Base		Achievement			Text version			Interaction		
Measure	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*	*p*	η_p^2	*F*	*p*	η_p^2	*F*	*p*	η_p^2
Final scores	76.1	12.5	63.5*	12.1	47.0	13.2	47.6	13.5	103.2	.0001	.71*	3.58	.065	.08*	4.28	.045	.09*
Initial scores	69.0	14.9	56.7*	15.6	41.7	14.7	45.1	15.6	50.7	.0001	.55*	1.46	.23	.03	4.56	.039	.10*
JCOMP	82.1	12.5	76.8	12.4	82.4	13.2	77.9	17.2	< 1	--	--	4.47	.040	.10*	< 1	--	--
Study time (min.)	5.5	1.7	4.5*	1.3	5.1	1.6	6.3*	2.1	2.95	.093	.07*	< 1	--	--	5.85	.020	.12*
Test time (min.)	7.3	2.3	7.6	3.0	7.8	4.1	5.2*	2.5	1.42	.240	.03	5.51	.024	.12*	9.06	.004	.18*
Reanswering time (min.)	6.4	2.5	5.4	2.4	5.6	2.7	5.5	2.7	< 1	--	--	1.14	.29	.03	< 1	--	--
Confidence	81.7	9.0	78.3	16.4	74.2	18.8	75.9	13.8	1.77	.191	.04	< 1	--	--	= 1	--	--
Overconfidence	12.7	18.2	21.6	12.3	32.5	16.5	30.8	13.7	17.50	.0001	.29*	1.41	.24	.03	3.03	.089	.07*
Resolution	.43	.56	.20	.54	.31	.53	.14	.43	< 1	--	--	4.28	.045	.09*	< 1	--	--
Control sensitivity	.89	.27	.86	.31	.72	.51	.56	.60	5.67	.024	.16*	< 1	--	--	< 1	--	--

* Effect size which is at least medium: Cohen's $d > 0.50$ for paired comparisons; $\eta_p^2 > .06$ for ANOVAs.

[1] For the time analyses it was $F(1, 41)$ because of missing values. For control sensitivity it was $F(1, 31)$.

To minimize the effects of general ability factors such as reading ability and working memory capacity on the MLR results, we also examined differences in participants' SAT scores between the groups and their association with the text versions. As expected, the lower achievers had lower SAT scores (M = 589, SD = 80.9) than the higher achievers (M = 666, SD = 74.6), $t(42) = 3.26$, $p < .01$, d = 1.01, and the correlation between SAT and test scores was significant, r =.51, $p <$ .0001. An ANCOVA, similar to the above ANOVA, with SAT as a covariate yielded highly similar results with a marginally significant effect of SAT with a medium effect size, $F(1, 41) = 3.39$, MSE = 101.94, p = .073, η_p^2 = .08. Importantly, the interactive effect of achievement and text version on test scores was not attenuated by the inclusion of SAT score as a covariate, $F(1, 41) = 4.67$, MSE = 225.07, p = .037, η_p^2 = .10. Additional analyses further suggest a lack of predictive value for the SAT score. Specifically, for each version, the correlation across participants between their test score and the score difference between the two versions was significant, r = .74, $p < .0001$ and r = .50, p = .001, for the concrete and base versions respectively. This suggests that the higher the overall score, the larger the advantage of the concrete over the base versions. Correlating this test score difference between the versions by SAT scores resulted in a very low correlation, r = .04. We conclude that the difference between the lower and higher achievers in terms of how concreteness affected performance at test is not attributable to the general ability reflected by SAT scores. Using other measures, we examined whether the difference in performance between the groups was associated with differences in MLR.

Judgment of comprehension (JCOMP)

Overall, the JCOMP ratings were inflated relative to the scores achieved in all tests (see Table 1), even though the test scores also reflect the extra processing that took place after eliciting the JCOMP. This finding alone is not very informative, because participants may have expected easier test questions, or used the scale without taking its numerical values to reflect anticipated performance at test. The important information lies in the differences between the groups and between the text versions.

The effect of the text version on JCOMP ratings was examined in relation to the achievement levels. An ANOVA, as above, on the JCOMP ratings revealed no effect of the achievement group. The only significant effect was the predicted main effect of the text version: Regardless of achievement level, the concrete versions gave rise to a higher JCOMP than the base versions.

Time regulation

Figure 1 presents graphically the amount of time devoted to each text version by lower and higher achievers in all three phases. Three participants did not provide the time for one of the phases; the analysis is based on all the available data. As can be seen in the figure, the total time devoted to the concrete versions (M = 18.8, SD = 3.7) was somewhat longer than the time invested in the base versions (M = 17.1, SD = 4.6), $t(40) = 1.77$, $p = .084$, $d = .28$. This difference accords perfectly with the 10% difference in text length, and thus the two can be accounted as equivalent with regard to MLR.

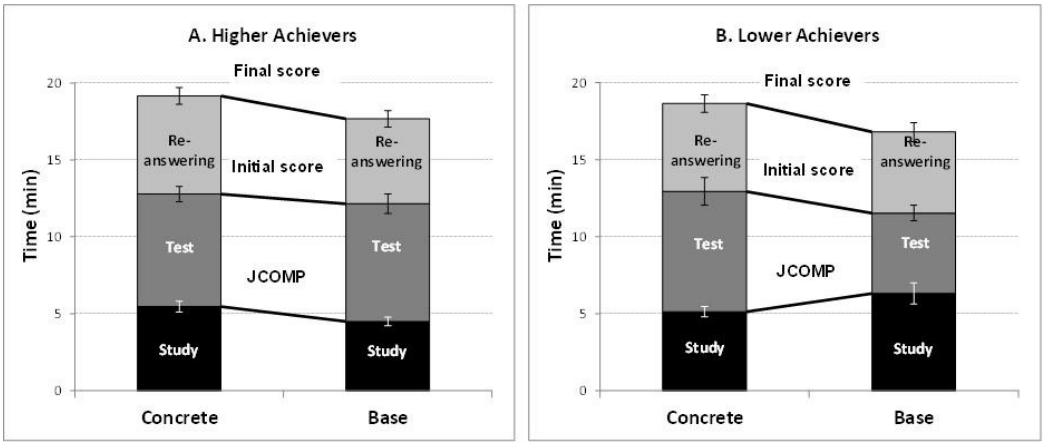

Figure 1. Time invested in concrete and base text versions by higher (panel A) and lower (panel B) achievers divided into the three phases: study, open-book test and reanswering two questions. The error bars represent standard errors of the mean. Time points for collecting the Judgment of Comprehension (JCOMP) and test scores are indicated.

Overall, it is evident from the figure that participants found it useful to invest time in each of the three phases. A three-way ANOVA for the effect of Achievement (low vs. high) × Text version (concrete vs. base) × Phase (study, initial test, and reanswering) on the time invested yielded a main effect of the phase, $F(2, 76) = 4.76$, $MSE = 9.76$, $p = .011$, $\eta_p^2 = .11$, and a significant triple interaction, $F(1, 76) = 7.90$, $MSE = 4.74$, $p = .001$, $\eta_p^2 = .17$. Because of the centrality of this finding, we also conducted a similar three-way ANCOVA with Achievement as a continuous covariate. With this analysis, the large effect size of the triple interaction became even stronger, $F(2, 76) = 10.27$, $MSE = 4.51$, $p < .0001$, $\eta_p^2 = .21$. Thus, as can be seen in Figure 1, although the overall time invested by the higher achievers was similar to that of the lower achievers, $F < 1$, it was divided differently among the phases.

To identify the source of the triple interaction, we performed a two-way ANOVA of Achievement × Text version for each of the three phases (see Table 1). There was a medium effect size for the achievement level and a significant interactive effect. Whereas the higher achievers studied the concrete versions longer, the lower achievers tended to invest less time studying these versions (with a medium effect size of 0.66). As explained above, this finding points to different strategic regulation during the study phase for the two groups. The lower achievers show an inverse relationship between JCOMP ratings and time investment, while the higher achievers show a positive relationship between them.

Examining the pattern of time investment during the initial test phase using ANOVA, as above, yielded no main effect of achievement but a main effect of text version accompanied by a large interactive effect. The higher achievers invested a similar amount of time in answering the test questions for the two versions, while the lower achievers invested significantly longer in taking the test related to the concrete versions. Thus, the lower achievers took advantage of the opportunity given to them during the test to regulate their answering differently for the two versions.

Finally, no differences were found between the groups or the versions in the time invested during the reanswering phase, all *p*s > .25. Thus, the processing differences generated by the inclusion of concrete supplements do not seem to extend through this later phase.

Confidence in test answers

There were no differences in mean confidence between the achievement groups or the text versions. Overall, calibration bias was in the direction of overconfidence.

Because of the large difference between the groups in performance in the initial test, which was not reflected in the confidence ratings, overconfidence was smaller for the higher than for the lower achievers. A two-way ANOVA, as above, with overconfidence as the dependent variable showed a main effect of achievement, no main effect of text version, and a marginal interactive effect which was medium in size (see Table 1).

Resolution of the confidence judgments was measured by computing for each participant the gamma correlation between confidence ratings and answer scores. A two-way ANOVA as above with resolution as dependent variable found no effect for the achievement group. However, over the two groups the mean correlation was higher for the concrete than for the base versions (see Table 1), and it was significantly higher than zero for both versions, $t(42) = 4.46$, $p < .0001$, $d = 0.68$ and $t(43) = 2.53$, $p = .015$, $d = 0.39$, respectively. Thus, the participants had a sense of which questions they had answered better than others for both versions, but this was especially marked in relation to the concrete versions. As mentioned above, stronger resolution may indicate deeper processing.

Control sensitivity

To examine the extent to which the participants followed their confidence ratings when choosing which questions to reanswer, the gamma correlation between confidence in a given answer and the choice to reanswer was calculated for each text studied by each participant. Because of the statistical limitations of gamma correlation (Spellman, Bloomfield, & Bjork, 2008), these data were available for 33 participants only. A two-way ANOVA revealed a significant difference between the higher ($N = 18$) and lower achievers ($N = 15$). The higher achievers

followed their confidence ratings very consistently and chose to reanswer the questions for which they were less confident about their answers. Lower achievers showed weaker control sensitivity. No effect was found for the text version or the interaction, probably because of the large variance, in particular in the lower achieving group.

Discussion

The present study aimed to shed light on how the inclusion of misleading surface-level cues in study materials affects MLR. For this purpose we added common concrete supplements in expository texts. The use of open-book testing served our aim to examine MLR along both study and test taking. Tracking metacognitive monitoring and self-regulation of learning efforts throughout allowed us to examine the association between MLR characteristics and performance at test.

We hypothesized that concreteness is used as a heuristic cue for JCOMP regarding expository texts, in a manner similar to that previously found with other tasks (Ackerman et al., 2013; Sadoski et al., 1993; Serra & Dunlosky, 2010). Indeed, the findings of the present study are similar to previous findings in terms of the higher JCOMPs and equivalent time investment in the presence of uninformative concrete supplements. However, unlike previous studies, which assumed universal effects, in the present study the partition into achievement groups and the integrative analysis of a comprehensive set of MLR measures enabled us to unravel the intricate relationships between metacognitive monitoring, regulation of time and choice, and ultimate performance. Notably, in line with our hypotheses, the universal effect of the supplements on JCOMP ratings appeared to modulate learning regulation differently for different people,

and this was associated with large score differences. Partialling out the predictive effects of SAT scores revealed that the found performance effects are pronounced above and beyond variability in general cognitive abilities. The various interactive effects indicate that the two groups differed in their sensitivity to the surface-level cue of concreteness. Importantly, we found a particularly pronounced difference – almost 30% – between the groups' test scores following study of the concrete versions, compared with 16% for the base versions (see Table 1).

Metacognitive regulation during the study phase

In terms of the distinction between data-driven and goal-driven allocation of study time and monitoring (Koriat et al. 2006), it seems that for the lower achievers, time investment was more data-driven, while for the higher achievers, it was more goal-driven. This suggests that the lower achievers were those who were misled more by the concreteness of the texts, whereas the higher achievers invested longer in the presence of the concrete supplements. As explained above, the inclusion of redundant graphical and textual content in texts was suggested to increase the cognitive load on learners, who must integrate the supplements with the text (Chandler & Sweller, 1991; Mayer et al., 2001), and this may take a toll on study time. Yet the higher achievers benefited from the inclusion of concrete supplements, despite their being uninformative.

How could these uninformative supplements improve the performance of the higher achievers? Reading comprehension involves a search after meaning (Graesser, Singer, & Trabasso, 1994). It is possible that the higher achievers expected the concrete supplements to be of some worth for understanding the text's contents, and this directed them to process the content in greater depth. This

explanation accords with the principle of *desirable difficulties* (Bjork, 1994), according to which learning situations which challenge learners lead them to invest more mental effort, and this deeper processing is beneficial for learning outcomes (see Bjork et al., 2013, for a review). Clearly, learning involves a delicate interplay between self-assessment, motivation, and effort (Pintrich, 2000). It may be that the higher achievers were those who were motivated to invest more effort in the search after meaning. This took longer and their effort investment had the benefit of improved performance for the concrete versions. Notably, our study used only uninformative supplements. Future studies are called for to examine the whether this benefit is greater when informative supplements are used, and whether their benefit is associated with MLR characteristics.

Another relevant consideration was raised recently by Miele, Finn, and Molden (2011). They found that the negative correlation often observed between time investment and metacognitive monitoring reflects data-driven effort investment only for individuals who believe that the task has brought them to the limit of their ability. In contrast, subjects who viewed intelligence as malleable, and who tended to interpret effortful encoding as indicating greater engagement in learning, predicted greater performance for items that they found more effortful to learn. This suggestion is similar to the goal-driven explanation of Koriat et al. (2006), but emphasizes individual differences in beliefs about intelligence. This idea accords well with our differential findings for higher and lower achievers, but a direct examination is called for.

Complementary metacognitive regulation during the test phase

In the present study, a comparison of the study and initial test phases reveals that the lower achievers invested somewhat less time in studying the concrete texts, but more time in taking the initial test for these versions. This finding suggests that when the lower achievers faced the test questions, they realized their initial learning was unsatisfactory, leading them to take advantage of the open-book procedure to compensate by investing additional effort during the test phase. This answers our question regarding the potential utilization of the opportunity encompassed in the open-book testing procedure. The extra processing during test taking allowed the participants to achieve equal performance for the two text versions. It also had the added benefit of improving resolution (discrimination between the better and more poorly answered questions). These findings point to effective strategic regulation of test taking as an action compensatory to study guided by what turned out to be a misleadingly high JCOMP.

For the higher achievers, the process seems to be different. Overall, this group was highly efficient, as they achieved much higher scores than the other group with a similar working time. They put more time into studying the concrete than the base versions. In the test phase, they invested a similar amount of time in the two versions, but their greater resolution suggests that they processed their answers for the concrete versions deeply and efficiently without investing additional time.

Finally, the higher achievers chose to reanswer the questions about which they were less confident even without investing extra time in test taking. This finding joins Ackerman and Goldsmith (2011) in showing variability of control sensitivity among young and healthy adults. In both studies, better control

sensitivity was associated with higher achievement. While the previous study examined variability in control sensitivity during the study phase, here we examined it during open-book testing. Future studies are invited to further investigate this important but understudied measure. In particular, our study used open-book testing so as to allow MLR to continue during the test phase, but clearly a comparison to closed-book testing is called for.

In sum, those who regulated their learning effectively throughout the process achieved higher scores. These participants worked efficiently, invested no more time than the lower achievers, and scored better both overall and for the concrete versions in particular. Importantly, partialling out the effects associated with SAT scores suggests that cognitive ability, verbal ability, exam taking skills, and other factors reflected by SAT scores do not underlie the differences between the achievement groups, and so justifies our focus on the quality of MLR.

Implications

Recognizing that concrete supplements can reduce the effectiveness of study by increasing the JCOMP even in the absence of improved understanding may help us interpret the results of past studies dealing with other types of understanding challenges. For example, Rozenblit and Keil (2002) found that the illusion of explanatory depth was greater for devices in which all components were visible, such as a bicycle, relative to devices in which the components were hidden, such as a cylinder lock. We would account for the difference in subjective feeling of understanding by noting that the JCOMP is similarly increased by the visibility of the components. Likewise, our results may be relevant to recent findings that participants judged explanations that contained irrelevant

neuroscience details or brain images as more convincing than the same explanations without those concrete details (McCabe & Castel, 2008; Weisberg, Keil, Goodstein, Rawson, & Gray, 2008). They may also account for the observation that highly detailed study aids, such as animated demonstrations used in computerized study and step-by-step tutoring, sometimes reduce students' ability to perform the relevant skill on their own or to transfer the acquired knowledge to new situations (e.g., Goldstone & Son, 2005; Lowe, 2004; Tversky, Morrison, & Betrancourt, 2002; Yuviler-Gavish, Yechiam, & Kallai, 2011). If such detailed and concrete study aids unduly increase the JCOMP, they may lead students to stop studying too early or to make poor study choices (e.g., failing to return to material they understand less well).

An important finding of the present study is that illustrations and concrete examples sometimes impede the study regulation process by unduly increasing learners' JCOMP, thereby reducing the effectiveness of MLR. Clearly our study is an initial examination, and we call for future studies to examine other heuristic cues that inform the JCOMP. It is also interesting to examine whether concreteness and perceived ease of processing (Serra & Dunlosky, 2010) converge. As an initial take on this question, pretest A revealed that the texts with the supplements were perceived as more concrete but not as less difficult. A parallel issue relates to the role of interest as a cue for JCOMP. Several studies have examined the effects of seductive details on learning. Seductive details are concrete details that make the text more interesting without contributing to an understanding of the main ideas conveyed in the text. A reduction in performance at test was found when the text included such details (Garner, Gillingham, &

White, 1989), and studies using seductive illustrations yielded similar results (Harp & Mayer, 1998). In the present study, Pretest A revealed that although the manipulated texts were judged to be more concrete, they were not judged to be more interesting. Future studies should examine the roles of the various potential cues for JCOMP and whether the previously found effects on performance are mediated by effects on JCOMP and regulation of effort.

Despite the misleading potential of concrete supplements, we certainly are not calling on the designers of study materials to exclude illustrations and concrete examples from textbooks, as they clearly can contribute to learning (Chen & Daehler, 2000; David, 1998; Sadoski, Goetz, & Avila, 1995). Rather, we join other researchers who emphasize that these supplements should be chosen with care, avoiding gratuitous illustrations and examples (Ainsworth, 2006; Carney & Levin, 2002; Levin & Mayer, 1993; Schnotz, 2005; Tversky et al., 2002; Vekiri, 2002). Indeed, while lower achievers are likely to suffer most from the inclusion of redundant supplements, some studies have found that when concrete supplements are carefully designed, students with low prior knowledge benefit the most from them (e.g., Mayer & Gallini, 1990).

Aside from the implications for the design of study materials, drawing attention to the potential risks posed by the presence of redundant supplements may encourage the development of guidelines for teachers, which would help them attenuate their students' vulnerability to uninformative supplements (e.g., diSessa, 2004; Rakes, 2008). Our findings also point to the possibility that teachers could use open-book testing, or other tools aimed at helping students engage in self-examination and improve their knowledge, to help learners

overcome the consequences of an inflated JCOMP. By posing these challenges, we join recent calls (de Bruin & van Gog, 2012; Son, 2007) for efforts to bridge the gap between the cognitive sciences and education by supporting the development of instruction programs focused on improving students' ability to reliably assess their knowledge and overcome factors that may mislead them.

References

Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. *Journal of Experimental Psychology: Applied, 17,* 18-32.

Ackerman, R., & Lauterman, T. (2012). Taking reading comprehension exams on screen or on paper? A metacognitive analysis of learning texts under time pressure. *Computers in Human Behavior*, *28*, 1816-1828.

Ackerman, R., Leiser, D., & Shpigelman, M. (2013). Is comprehension of problem solutions resistant against misleading heuristic cues? *Acta Psychologica*, *143*(1), 105-112.

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger III, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*, 861-876.

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction, 16*(3), 183-198.

Anderson, M. C. M., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica*, *128*, 110-118.

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

Bol, L., & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *The Journal of experimental education*, *69*(2), 133-151.

Bjork, 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). Cambridge, MA: MIT Press.

Bjork, R. A., Dunloksy, J., & Kornell, N. (2013). Self-Regulated Learning: Beliefs, Techniques, and Illusions. *Annual Review of Psychology, 64,* 417-444.

Brown, A. L. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. E. Weinert & R. H. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 65-116). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Carney, R. N., & Levin, J. R. (2002). Pictorial illustrations still improve students' learning from text. *Educational Psychology Review*, *14*, 5-26.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*(4), 293-332.

Chen, Z., & Daehler, M. W. (2000). External and internal instantiation of abstract information facilitates transfer in insight problem solving. *Contemporary Educational Psychology*, *25*, 423-449.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

David, P. (1998). News concreteness and visual-verbal association: Do news pictures narrow the recall gap between concrete and abstract news? *Human Communication Research*, *25*, 180-201.diSessa, A. A. (2004). Metarepresentation: Native competence and targets for instruction. *Cognition and Instruction, 22*, 293-331.

de Bruin, A. B. H., & van Gog, T. (2012). Improving self-monitoring and self-regulation: From cognitive psychology to the classroom. *Learning and Instruction*, 22, 245-252.

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

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist, 46*, 6-25.

Feller, M. (1994). Open-book testing and education for the future. *Studies in Educational Evaluation, 20*(2), 235–238.

Fisher, S. L., & Ford, J. K. (1998). Differential effects of learner effort and goal orientation on two learning outcomes. *Personnel Psychology, 51*(2), 397-420

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2.

Garner, R., Gillingham, M., & White, C. (1989). Effects of "seductive details" on macroprocessing and microprocessing in adults and children. *Cognition and Instruction*, 6, 41-57.

Goldstone, R. L., & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *Journal of the Learning Sciences*, *14*, 69-110.

Grant, H., & Dweck, C. S. (2003). Clarifying achievement goals and their impact. *Journal of Personality and Social Psychology, 85*(3), 541-553.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371-395.

Greene, J. A., & Azevedo, R. (2007). A Theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research*, *77*, 334.

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition, 36*(1), 93-103.

Hacker, D. J. (1998). Self-regulated comprehension during normal reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in education theory and practice* (pp. 165-192). Mahwah, NJ: Lawrence Erlbaum Associates.

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning, 3*(2), 101-121.

Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of Educational Psychology*, *90*, 414-434.

Heijne-Penninga, M., Kuks, J. B. M., Hofman, W. H. A., & Cohen-Schotanus, J. (2011). Directing students to profound open-book test preparation: The relationship between deep learning and open-book test time. *Medical Teacher, 33*(1), e16-e21.

Koren, D., Seidman, L. J., Poyurovsky, M., Goldsmith, M., Viksman, P., Zichel, S., & Klein, E. (2004). The neuropsychological basis of insight in first-episode schizophrenia: A pilot metacognitive study. *Schizophrenia research, 70,* 195-202.

Koriat, A., Ackerman, R., Adiv, S., Lockl, K., & Schneider, W. (in press). The effects of goal-driven and data-driven regulation on metacognitive monitoring during learning: A developmental perspective. *Journal of Experimental Psychology: General.*

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*, 490-517.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135*(1), 36-69.

Levin, J. R., & Mayer, R. E. (1993). Understanding illustrations in text. In B. K. Britton, A. Woodward, & M. Brinkley (Eds.), *Learning from textbooks: Theory and practice* (pp. 95–113). Hillsdale, NJ: Erlbaum.

Lowe, R. (2004). Interrogation of a dynamic visualization during learning. *Learning and Instruction*, *14*(3), 257-274.

Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker (Ed.), *Metacognition in Educational Theory and Practice. The Educational Psychology Series* (pp. 117-144). Mahwah, NJ, US: Lawrence Erlbaum Associates.

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 663-679.

Maki, R. H., Foley, J. M., Kajer, W. K., Thompson, R. C., & Willert, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *16*, 609-616.

Mayer, R. E. (2003). The promise of multimedia learning: using the same instructional design methods across different media. *Learning and Instruction, 13*(2), 125-139.

Mayer, R. E., & Gallini, J. K. (1990). When is an illustration worth ten thousand words? *Journal of Educational Psychology, 82*, 715-726.

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology, 93*(1), 187-198.

McCabe, D. P., & Castel, A. D. (2008). Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition*, *107*, 343-352.

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.

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

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

Novick, L. R., & Sherman, S. J. (2008). The effects of superficial and structural information on online problem solving for good versus poor anagram solvers. *The Quarterly Journal of Experimental Psychology, 61*, 1098–1120.

Pansky, A., Koriat, A., Goldsmith, M., & Pearlman-Avnion, S. (2009). Memory accuracy and distortion in old age: Cognitive, metacognitive, and neurocognitive determinants. *European Journal of Cognitive Psychology, 21,* 303-329.

Pintrich, P. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. Pintrich & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). San Diego, CA: Academic Press.

Prangsma, M. E., Boxtel, C. A. M., Kanselaar, G., & Kirschner, P. A. (2009). Concrete and abstract visualizations in history learning tasks. *British Journal of Educational Psychology, 79*(2), 371-387.

Pressley, M. (2002). Metacognition and self-regulated comprehension. In A. E. Farstrup & S. J. Samuels (Eds.), *What research has to say about reading*

instruction (3 ed., pp. 291-309). Newark, DE: International Reading Association.

Rakes, G. C. (2008). Open book testing in online learning environments. *Journal of Interactive Online Learning, 7*(1), 1-9.

Richardson, J. T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*(2), 135-147.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*, 521-562.

Sadoski, M., Goetz, E. T., & Avila, E. (1995). Concreteness effects in text recall: Dual coding or context availability? *Reading Research Quarterly*, *30*, 278-288.

Sadoski, M., Goetz, E. T., & Fritz, J. B. (1993). Impact of concreteness on comprehensibility, interest, and memory for text: Implications for dual coding theory and text design. *Journal of Educational Psychology*, *85*, 291-304.

Schnotz, W. (2005). An integrated model of text and picture comprehension. In R. E. Mayer (Eds.), *The Cambridge handbook of multimedia learning* (pp. 49-69). Cambridge, MA: Cambridge Univ. Press.

Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgements reflect the belief that diagrams improve learning from text. *Memory*, *18*, 698-711.

diSessa, A. A. (2004). Metarepresentation: Native competence and targets for instruction. *Cognition and Instruction*, 22(3), 293-331.

Son, L. K. (2007). Introduction: A metacognition bridge. *The European Journal of Cognitive Psychology*, *19*, 481-493.

Spellman, B. A., Bloomfield, A., & Bjork, R. A. (2008). Measuring memory and metamemory: Theoretical and statistical problems with assessing learning (in general) and using gamma (in particular) to do so. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of Metamemory and Memory* (pp. 95-114). New York: Psychology Press.

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

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *31*, 1267-1280.

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

Thiede, K. W., Wiley, J., & Griffin, T. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology, 81,* 264-273.

Tversky, B., Morrison, J. B., & Betrancourt, M. (2002). Animation: Can it facilitate? *International Journal of Human-Computer Studies*, *57*, 247-262.

Vekiri, I. (2002). What is the value of graphical displays in learning? *Educational Psychology Review*, *14*, 261-312.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, *20*, 470-477.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker (Ed.), *Metacognition in educational theory and practice* (pp. 277-304). Mahwah, NJ, US: Lawrence Erlbaum Associates, Inc.

Yuviler-Gavish, N., Yechiam, E., & Kallai, A. (2011). Learning in multi-modal training: Visual guidance can be both appealing and disadvantageous in spatial tasks. *International Journal of Human-Computer Studies, 69*, 113-122.

Zaromb, F. M., Karpicke, J. D., & Roediger III, H. L. (2010). Comprehension as a basis for metacognitive judgments: Effects of effort after meaning on recall and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(2), 552-557.

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

Zohar A., & Dori, Y. J. (2012). *Metacognition in science education*. Springer, NY.

APPENDIX - An Example of Materials

Below we present an example of the study materials and manipulations used for this research. The underlined words and the drawings and pictures (in color) were added to the base text to generate the concrete version. The study materials were in Hebrew.

The nutritional value of spinach

Spinach leaves

Spinach has high nutritional value. It is extremely rich in antioxidants, especially when fresh, steamed, or quickly boiled. Spinach is considered a rich source of iron and calcium. However, the bioavailability of iron is dependent on its absorption. This is influenced by a number of factors.

Iron enters the body in two forms: non-heme iron and heme iron. The distinction is that certain foods and drinks, such as those that include caffeine like tea and coffee, can interfere with the absorption of non-heme iron, whereas heme iron is digested even with their presence. All of the iron in grains and vegetables, and about three-fifths of the iron in animal food sources (meats), is non-heme iron. The much smaller remaining portion from meats is heme iron.

Popeye

Popeye the Sailor Man has a strong fondness for spinach, becoming physically stronger after consuming it. However, the iron in spinach is poorly absorbed by the body unless eaten with vitamin C. The type of iron found in spinach is non-blood (non-heme), a plant iron, which the body does not absorb as efficiently as blood (heme) iron, found in meat.

Examples of test questions:

1. Ron is about to prepare lunch for his family, and he would like to use the opportunity to maximize the iron contribution to his family's nutrition. He would like to prepare a spinach dish. Suggest for Ron an appropriate dish with detailed instructions for him to follow.
2. It is well known that Coke includes caffeine. Given that Ron's family is particularly fond of Coke during lunchtime, what would you advise Ron to serve as the main dish alongside his spinach dish: baked whole-grain rice with almonds, or meatballs? Explain your recommendation.