Reference:
Lauterman, T., & Ackerman, R. (in press). Overcoming screen inferiority in learning and calibration. *Computers in Human Behavior.*

January, 2014

Overcoming Screen Inferiority in Learning and Calibration

Tirza Lauterman and Rakefet Ackerman

Faculty of Industrial Engineering and Management,

Technion–Israel Institute of Technology, Haifa, Israel

E-mail addresses: Tirzal@tx.technion.ac.il; Ackerman@ie.technion.ac.il

Author note: The study was supported by a grant from the Israel Science Foundation (Grant No. 957/13). Partial data (N = 80 out of 163) with a reduced scope of analyses were reported in the proceedings of the 35th annual conference of the Cognitive Science Society (Lauterman & Ackerman, 2013). We are grateful to Meira Ben-Gad for editorial assistance.

Abstract

Metacognitive monitoring that accompanies a learning task reflects self-prediction of achievement at test. Well-calibrated monitoring is important because it is by this subjective assessment that people allocate their learning efforts. Previous studies that compared learning outcomes and calibration of monitoring when learning texts on screen and on paper have found screen inferiority: screen learners performed worse and were more overconfident about their success. However, learning from one's preferred medium was associated with attenuated overconfidence. The present study examined two methods for overcoming screen inferiority in these respects. First, practicing the study-test task allowed overcoming screen inferiority, but only among those who preferred reading from screens. Second, in-depth processing was encouraged by having participants generate keywords at a delay, before monitoring their knowledge and taking the test. This method eliminated screen inferiority even for the first-studied texts, but after practicing it, screen inferiority was re-exposed among those who preferred studying on paper. This study makes a practical contribution to educational practice by suggesting directions for overcoming screen inferiority. From a broader perspective, the study demonstrates that experience with the task and in-depth processing can attenuate overconfidence and that the effectiveness of learning-enhancing methods depends on the study context and learners' preferences.

Keywords: Reading comprehension; e-learning; human-computer interaction; metacognitive monitoring; metacomprehension; overconfidence.

1. Introduction

Theories of self-regulated learning suggest that spontaneous subjective assessment of knowledge, or metacognitive monitoring, plays an important role in learning regulation (Nelson & Narens, 1990), in addition to the conscious use of learning strategies and assessment of their effectiveness (see Winne & Hadwin, 1998, and Greene & Azevedo, 2007, for reviews). Indeed, empirical studies dealing with memorization and reading comprehension tasks have shown an association between monitoring output and decisions regarding allocation of study time (Metcalfe & Finn, 2008; Thiede, Anderson & Therriault, 2003). However, studies dealing with reading comprehension tasks have found that the accuracy of the relevant metacognitive judgments – metacomprehension judgment or prediction of performance at test – tends to be particularly poor (Dunlosky & Lipko, 2007; Thiede, Griffin, Wiley & Redford, 2009).

It is well established that metacognitive monitoring is not always reliable, and that this is because learners base their judgments on heuristic cues (Koriat, 1997; see Bjork, Dunsloky, & Kornell, 2013, and Dunlosky & Tauber, in press, for reviews). Although this theory was originally developed in the context of memorization tasks, a body of research has suggested that such cues are similarly used to judge comprehension. Cues found to take part in metacomprehension judgments include domain familiarity and interest in the topic (Glenberg & Epstein, 1987; Maki & Serra, 1992), accessibility of information in memory (Baker & Dunlosky, 2006), text concreteness (Ackerman & Leiser, in press), ease of processing the text (Dunlosky & Rawson, 2005; Maki, Foley, Kajer, Thompson, & Willert, 1990; Rawson & Dunlosky, 2002), and global characteristics of texts such as length or difficulty (Weaver & Bryant, 1995). According

to this literature, the accuracy of metacomprehension judgments is affected by the predictive validity of these and other cues used in the metacomprehension process.

Previous studies which examined factors that affect metacomprehension accuracy dealt, in the main, with characteristics of the learners (e.g., Griffin, Wiley, & Thiede, 2008), the particulars of the tasks (e.g., Thiede et al., 2003), or characteristics of the text's contents or design (e.g., Ackerman, Leiser, & Shpigelman, 2013; Rawson & Dunlosky, 2002). The present study broadens this inquiry in line with theories highlighting that learners' beliefs regarding the effectiveness of computer-supported learning environments modulate the ways in which these learning environments are used, the goals people set for their learning, and the expected outcomes (e.g., Antonietti & Colombo, 2008). Examinations of these theories often focus on conditions, involving both individual differences and design of the learning task, that enable effective utilization of unique features found in computerized learning environments but not on paper, like multimedia and hypertext (e.g., Antonietti , Colombo, & Lozotsev, 2008; Azevedo, 2005; Veenman, Prins, & Elshout, 2002).

We took into account individual differences in beliefs regarding the effectiveness of learning on screen versus on paper, by extending a line of research analyzing reading comprehension that can be performed comparably on both media (Ackerman & Goldsmith, 2011; Ackerman & Lauterman, 2012). In particular, it was found that peoples' medium preference affect their metacognitive processes when learning from texts (Ackerman & Lauterman, 2012). From a theoretical perspective, this approach enables a focus on how the medium through which learning takes place affects self-regulated learning by ruling out differences resulting from population, content, and

design-based characteristics. From a practical perspective, learning from continuous texts is widespread in computerized environments, and in many cases these environments offer no special features that are not found on paper. For instance, digital media provide an assortment of on-demand textual information for developing professional competence, like providing access to user reference books or academic papers. Students face computerized reading comprehension tasks in their studies, and higher education candidates face them in online screening exams (e.g., the Graduate Management Admission Test, the GMAT). Thus, it is important to consider the ways self-regulated learning is affected by the medium on which one learns.

Several studies have found screen inferiority in subjective and objective learning measures, as detailed below. In the present study, we aimed to offer methods for overcoming screen inferiority, while considering the study medium, screen versus paper, and participants' medium preference as factors in this improvement.

1.1 The effect of the study medium on text learning

There is growing evidence for cognitive and behavioral differences associated with learning from texts presented on screen and on paper. For example, students scored lower in reading comprehension tests after reading a text presented on screen compared with paper (Mangen, Walgermo, & Brønnick, 2013). Liu (2005) found in a self-report study that when reading on screen, people tend to engage more in browsing and scanning, one-time reading, and non-linear reading, with less sustained attention and less time spent reading in depth. Such findings suggest that people perceive reading from a screen as appropriate for a superficial kind of reading. Indeed, Morineau, Blanche, Tobin and Guéguen (2005) found that the mere presence of an e-book near a learner hindered recall

of information, while the presence of the paper book facilitated it. They suggested that the medium on which a text is presented provides a contextual cue for the retrieval process. It is possible that because of this perception, fewer cognitive resources are mobilized for the comprehension and metacomprehension processes when learning from a screen.

Only a few studies have examined the effects of the reading medium on metacomprehension processes. Ackerman and Goldsmith (2011) compared metacognitive monitoring and control during on-screen and paper learning when both groups of participants faced identical tasks. They measured participants' calibration bias–a measure of over- or underconfidence–by calculating the gap between the participants' mean Predictions of Performance (POPs) and test scores. On-screen learners (OSLs) showed more pronounced overconfidence than on-paper learners (OPLs). In accordance with their biased monitoring, OSLs studied the texts for a somewhat shorter time and achieved lower test scores than OPLs. Considering the increasing prevalence of on-screen learning, it is worth looking into methods for overcoming the screen inferiority found with respect to both performance and calibration bias.

Overall, people tend to prefer reading texts in depth from print rather than from computerized environments (Buzzetto-More, Sweat-Guy, & Elobaid, 2007; Jamali, Nicholas, & Rowlands, 2009; Spencer, 2006; Woody, Daniel, & Baker, 2010). Indeed, the screen inferiority found by Ackerman and Goldsmith (2011) was obtained from students who strongly preferred print over computerized learning. Using students who had only a moderate preference for print, Ackerman and Lauterman (2012) found similar screen inferiority only under mild time pressure. On the one hand, this finding suggests

that these students could overcome screen inferiority when they did not have the additional burden of adhering to a time limit. On the other hand, this finding also suggests that screen inferiority remains potent even among learners who have a more positive attitude towards this study medium. Interestingly, Ackerman and Lauterman (2012) found that the best calibration was achieved in both media by those who studied on their preferred medium. Thus, learners' preference seems to be an additional important factor in the accuracy of knowledge monitoring and in the effectiveness of learning regulation according to task demands even for learning from continuous texts.

1.2 Metacomprehension improvement

Metacomprehension research combines reading comprehension theories with metacognition theories. Kintsch (1998) proposed a model of representation levels to explain the processes involved in reading comprehension. According to this model, readers construct meaning from a text at three levels: surface level – the information conveyed by words and signs; the relationships between words that comprise sentences; and at the highest level, the extraction of meaning not conveyed directly by the words and their relationships, a process that Kintsch calls inference or situational representation. It can be derived from this theory that when high-order comprehension is tested, POP should be more accurate when it relies on cues related to high-level representation of the text.

Indeed, two kinds of manipulations aimed at improving high-level representation have been shown to enhance participants' monitoring accuracy. The first of these is practice with the task and test. Hacker, Bol, Horgan, and Rakow (2000) found, in a natural classroom setting, better calibration of POPs as students gained more test practice

during the course. Attenuation of overconfidence with practice, and even underconfidence, were found in memorization tasks (e.g., Koriat, Sheffer, & Ma'ayan, 2002). The present study examined whether overconfidence would also be reduced when participants practiced text learning and subsequent test-taking over one session.

The second approach shown to improve monitoring accuracy involves encouraging learners to engage in in-depth processing of the studied text. For example, asking participants to generate keywords or to write a summary of the text after a delay consistently improved monitoring accuracy (Anderson & Thiede, 2008; Fukaya, in press; Thiede, Dunlosky, Griffin, & Wiley, 2005). In another study, improvement was achieved by instilling test expectancy directed to the level of processing required by the test (Thiede, Wiley, & Griffin, 2011). These methods for enhancing depth of processing proved effective for improving resolution – that is, the extent to which metacognitive judgments discriminate between better- and lesser-known items. Thiede and his colleagues did not examine calibration bias. The present study examined whether such in-depth processing methods are also effective for attenuating overconfidence.

Notably, many of the studies that have found improvements in monitoring accuracy were conducted in computerized environments (e.g., Anderson & Thiede, 2008). The present study examined whether such methods are particularly effective on screen, where processing is hypothesized to be shallower. Examining this hypothesis is important in two respects. First, it may point to practical opportunities for attenuating screen inferiority. Second, it has theoretical significance in showing that the extent of improvement depends on study context, beyond variables related to the learners and/or the task.

1.3 Overview of the study

The present study examined two approaches – (1) practice and (2) increasing participants' depth of processing – that have been found to enhance learning and monitoring accuracy in other settings. We examined whether the study medium is associated with any difference in the effectiveness of these approaches. In particular, we hypothesized that with both approaches, performance and calibration would improve more for on-screen learners than for paper learners, who process texts more deeply by default. We also hypothesized that any improvements found would be greater for participants learning on their preferred medium.

2. Experiment 1

As explained above, previous results suggest that practice may help in attenuating overconfidence, beyond its effect on performance at test (Hacker et al., 2000; Koriat et al., 2002). The present study examined whether practice improves performance and attenuates overconfidence in text learning within one session, and whether the study medium affects this improvement.

Overall, we adapted and extended the methodology used by Ackerman and Lauterman (2012, Experiment 1). In particular, we used the same population of engineering students, which is characterized by only a moderate paper preference, with an almost equal percentage of students preferring learning on screen and on paper. With this population, Ackerman and Lauterman found that performance and overconfidence on screen were inferior to paper only under mild time pressure.

As for test expectancy (Thiede et al., 2011), like Ackerman and Lauterman (2012), we informed the participants in advance that the test questions would examine

both memory for details and higher-order comprehension. In Ackerman and Lauterman's (2012) study, each participant studied and was tested on two texts in each condition. In the present study, we provided the participants with an opportunity to practice by performing the same study-test procedure with six texts. No feedback was provided. We hypothesized that practice would allow the participants to adjust their depth of processing to the test requirements and would result in improved test scores and attenuated calibration bias from the first to the final tests. Importantly, we expected these improvements to be more pronounced on screen than on paper, and more pronounced for the participants who studied from their preferred medium.

2.1 Method

2.1.1 Participants. Eighty-seven undergraduate engineering students (M_{age} = 25 years, 50% females) participated in the study for course credit. All participants reported having no learning disabilities. They were randomly assigned to one of two groups identified by the medium, On-Screen Learners (OSLs; N = 38) or On-Paper Learners (OPLs; N = 49).

2.1.2 Materials. The six texts, 1000-1200 words (2-4 pages) each, dealt with various topics (e.g., the advantages of coal-based power compared to other energy sources; adult initiation ceremonies in various cultures). An additional, shorter, text (200 words) was used to familiarize the participants with the procedure. The texts were taken from websites intended for reading on screen. The texts were randomly assigned to their position in the study list for each participant. Each text formed the basis for a multiple-choice test comprising five questions testing memory of details and five questions testing higher-order comprehension.

A printed self-report questionnaire asked for a few personal details, including SAT scores. The critical questions asked about the respondent's generally preferred study medium, paper or screen, when studying thoroughly a text found on the web or received by e-mail; and about the medium perceived as producing more effective learning – paper, screen, or no difference.

2.1.3 Procedure. The experiment was administered in groups of up to eight participants in a small computer lab. All participants in each group worked on the same medium. Each participant studied and was tested on six texts. The procedure for each text included consecutive study, POP, and test phases, in a manner identical to that used by Ackerman and Lauterman (2012), as detailed below. At the start, the participants read the general instructions from a printed booklet. They were told that they would be asked to study for a multiple-choice test that would assess both their memory for details and their higher-order comprehension. The instructions for the OSL groups included explicit permission to edit the file. They also included guidance regarding Microsoft Word annotation tools (boldface, underlining, highlighting, font color, and marginal comments), although participants were familiar with these tools beforehand. Paper participants were provided with a yellow marker and a pen for note-taking.

For both media, the experiment was administered by a computer program. For the OSL groups, when the "Start" button was pressed, the program opened the relevant text in Microsoft Word, in edit mode. When participants finished learning, they saved the file and closed the program, then pressed the "Continue" button on the screen. For the OPL groups, the six texts were presented face-down in a pile at each station. Pressing the "Start" button opened a window on the screen which directed participants to turn over the

next text to be studied. The participants then took the printed text from the top of the pile and began reading. When they finished learning, they turned the text face-down again and pressed the "Continue" button.

Two predictions of performance (POPs) were collected on the screen immediately after the participants studied each text. Participants were asked to drag an arrow along a continuous 25%-100% scale to indicate how well they thought they would do on test questions that involved (1) memory for details and (2) higher-order comprehension.

The procedure for the test was similar to that of the study phase. For the OSL group, the test form opened in Microsoft Word. Participants marked their chosen option for each question using the highlighter tool and saved the file. The OPL group marked their answers using a yellow marker on paper.

The experiment began with a run of the entire task (study, POPs, and test) on the target medium, using the shorter practice text. Then the participants were informed that they would be given 7 minutes to study each text and 5 minutes for each of the six tests. The study time was intentionally short, as participants in previous studies took 9-10 minutes on average when they were free to regulate their study time. The test time, in contrast, was sufficient to allow unpressured answering. The participants were informed when there was one minute left in each phase.

The participants filled in the self-report questionnaire after completing the experimental procedure. The whole procedure, including instructions and the practice text, took about 90 minutes.

2.2. Results

Test scores were calculated as the percentage of correctly answered questions out of ten. The scores were significantly higher than chance level (25%) and lower than perfect performance (100%) for each of the six texts, all *p*s < .0001, assuring sufficient variability and allowing enough margin for POPs to show under- or overconfidence. The mean SAT score of the sample was 678.2 (*SD* = 37.4),[1] with no difference between the media groups, $t < 1$.

In the self-report, 61% of the participants stated that they would print the text for thorough learning. With regard to study effectiveness, 56% responded that paper would produce better outcomes, 5% ($N = 4$) responded likewise for screen learning, and 39% expected no effect of the medium. Importantly, no difference in medium preference was found between the OSL and OPL groups, $\chi^2(1) = 0.004$, $p = .95$.

To examine the cumulative effect of practice, the dependent variables were averaged across three text pairs: the first and second texts (Pair 1), the third and fourth (Pair 2), and the fifth and sixth (Pair 3). The same procedure was followed for POPs and calibration bias. Participants were rated as overconfident when their mean POP for the two texts in a given pair was higher than their mean test score for those texts.

2.2.1 Verification of screen inferiority

We first examined the effect of the medium on Pair 1, in order to verify that the screen inferiority under time pressure found by Ackerman and Lauterman (2012) was replicated here. Fig. 1 presents the results.

[1] Israeli SAT scores range from 200-800; the national mean is 540; 15% of test-takers nationwide score above 650; 5% score above 700.

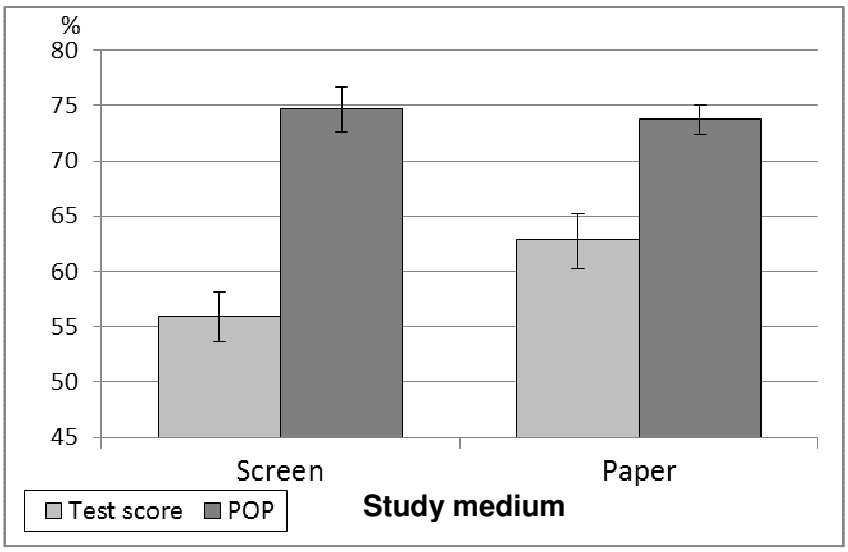

Fig. 1. Mean test scores and predictions of performance (POP) for Pair 1 on screen and on paper. Error bars represent the standard errors of the means.

To rule out potential effects of general ability on test scores, we used SAT scores as a covariate in all analyses that involved test scores. In all cases, there was a main effect of the SAT scores, all *p*s < .05, but this effect is beyond the scope of this study. A two-way Analysis of Covariance (ANCOVA) examining the effects of Measure (POP or score) × Medium (screen or paper) on test scores yielded a main effect of the measure, $F(1, 82) = 34.47$, $p < .0001$, $\eta_p^2 = .30$, with both groups showing overconfidence, both *p*s < .0001 for the differences between POP and test scores. However, the interactive effect was also significant, $F(1, 82) = 6.36$, $p = .01$, $\eta_p^2 = .07$. As can be seen in Fig. 1, OSLs scored lower than OPLs, $t(85) = 2.21$, $p < .05$, but showed no difference in POPs, $t < 1$. Thus, the interaction stemmed from more pronounced overconfidence for OSLs than for OPLs. These findings replicate the results of Ackerman and Lauterman (2012) and provide a starting point for the present study.

2.2.2 The effect of practice

The effect of practice was examined by calculating the same ANCOVA described above for Pairs 2 and 3. For Pair 2, we again found a main effect of the measure, $F(1, 82) = 8.84$, $p < .005$, $\eta_p^2 = .10$, with both OSLs and OPLs showing overconfidence, both ps < .01 for the difference between POP and test scores. But this time, there was no interactive effect, $F < 1$, meaning that OSLs and OPLs showed similar levels of overconfidence. Pair 3 resembled Pair 2, with a main effect of the measure, $F(1, 82) = 13.15$, $p < .001$, $\eta_p^2 = .14$, and no interaction. Thus, even a little practice neutralized the screen inferiority found in Pair 1.

In light of our predictions about the role of medium preference, we examined for each medium the combined effect of practice and preferred medium on test scores. In Fig. 2, the two panels in each row represent the same medium, screen (Panel A and Panel B) or paper (Panel C and Panel D), while the two panels in each column represent learning on the preferred (Panel A and Panel C) or non-preferred medium (Panel B and Panel D). A two-way ANCOVA of Practice (pair 1, 2, 3) × Preferred Medium (screen vs. paper) on test scores for OSLs yielded a significant interactive effect, $F(2, 68) = 5.32$, $p < .01$, $\eta_p^2 = .14$. OSLs who studied from their preferred medium (Panel A) improved their test scores by about 18 (!) points from Pair 1 to Pair 3, $t(14) = 5.45$, $p < .001$, while OSLs who preferred studying from paper (Panel B) did not improve their test scores, $t < 1$. The same examination for OPLs (Panel C and Panel D) showed no improvement from practicing the task and no effect of preference, all Fs < 1. Overall, OSLs who studied from their preferred medium caught up with OPLs, achieving equally good test scores by

Pair 3, $p = .25$. These scores were also better than those of OSLs preferring paper at a level close to significance, $t(36) = 1.91$, $p = .06$.

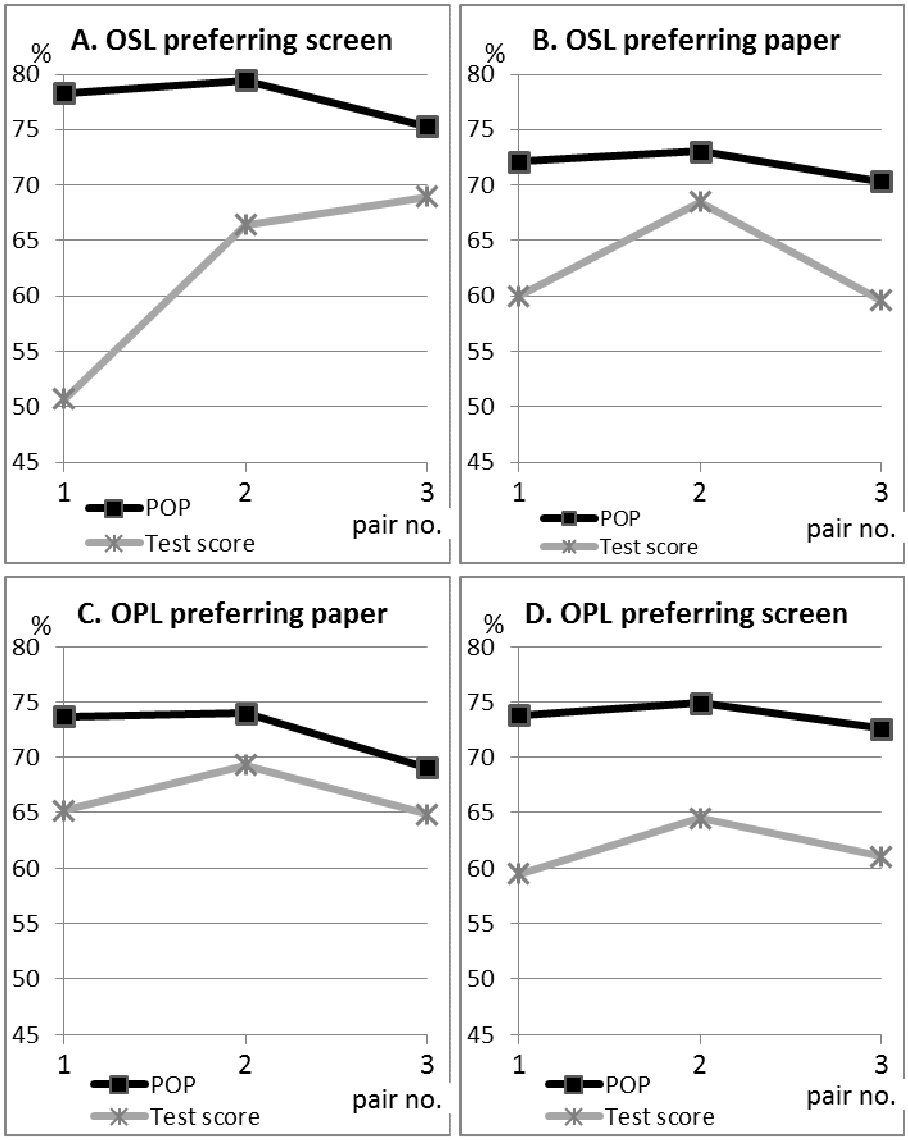

Fig. 2. Mean test scores and predictions of performance (POP) for the three text pairs by their study order, in division by study medium (On-Paper Learners – OPL, On-Screen Learners – OSL) and medium preference.

We next examined whether the observed patterns in test scores were reflected in participants' POPs. A two-way ANOVA of Practice (pair 1, 2, 3) × Preferred Medium

(screen vs. paper) on POP for OSL (Panels A and B) yielded no main effects and no interactive effect, *F*s < 1. The same ANOVA for OPL (Panels C and D) yielded a significant main effect of practice, $F(2, 94) = 3.52$, $p < .05$, $\eta_p^2 = .07$, with more modest POPs after practice, $t(48) = 2.09$, $p < .05$ for the difference between Pair 1 and 3. The effects of practice on POPs and test scores led to changes in the extent of overconfidence. A three-way ANCOVA of Practice (pair 1, 2, 3) × Medium (screen vs. paper) × Preferred Medium (screen vs. paper) on overconfidence yielded a significant triple interaction, $F(2, 160) = 4.23$, $p < .05$, $\eta_p^2 = .05$. Among OSLs and OPLs who studied from their preferred medium, overconfidence was attenuated with practice, while those who studied from their non-preferred medium remained at the same level of overconfidence. Among OSLs who studied from their preferred medium, the difference between Pair 1 and Pair 3 was significant, $t(13) = 6.06$, $p < .001$. For OPLs who studied from their preferred medium, overconfidence was attenuated only marginally, $p = .1$, but it became well-calibrated (overconfidence not different from zero, $p = .20$).

2.3. Discussion

The results of the first text pair replicated the previously found screen inferiority (Ackerman & Lauterman, 2012), with lower performance and greater overconfidence for screen compared with paper learners. The present study examined whether practicing the learning task by repeating the task with different texts would allow learners to overcome this inferiority. This was indeed found to be the case for OSLs who studied from their preferred medium. This group improved their test scores for the second and third pairs, and achieved scores equivalent to those of OPLs. The POPs of screen learners did not reflect the changes in test scores, meaning that for those who preferred learning on

screen, this resulted also in attenuation of their overconfidence. It is interesting to note that OPLs who studied from their preferred medium benefitted from repeating the same task as well, by adapting their POPs to actual performance at test. In both media, participants learning from their non-preferred medium did not benefit from practicing the task. Thus, repeating the same learning task did improve the performance and monitoring accuracy of screen learners, but only for those preferring reading from a screen.

3. Experiment 2

The purpose of Experiment 2 was to examine whether encouraging in-depth processing would eliminate screen inferiority from the beginning, even before participants practiced the task. For this purpose, we used delayed keyword generation, a method found to be effective in improving resolution (Thiede et al., 2003, 2005). Participants first studied two texts in a row. They then followed the following procedure for each text, in the order studied: they wrote down four keywords summarizing the text's essence, predicted their performance at test, then took the test. The delay in testing was expected to lower test scores relative to the immediate test in Experiment 1, because the passage of time and interference of the intermediate task were expected to cause some forgetfulness, although in-depth processing might attenuate this score reduction. With regard to POPs, monitoring immediately after learning, as done in Experiment 1, can be performed using the highly accessible surface representation achieved while studying but which may not be as accessible while taking the test. After a delay, the situation model (Kintsch, 1998) is more likely to be accessed and provide a more reliable basis for predicting performance at test (Thiede et al., 2003). We hypothesized that this would attenuate overconfidence, as Thiede et al. (2003, 2005) found for resolution.

What should we expect as to the effect of the medium on outcomes from this in-depth processing? In Experiment 1, it appears that OPLs who studied on their preferred medium did not need to practice: they achieved with Pair 1 scores as high as those which OSLs who studied from their preferred medium achieved only in Pair 3, and they showed the best calibration all along. We interpret these findings to suggest that the OPLs showed appropriate learning naturally, and by practicing they were able to fine-tune their monitoring, while the OSLs had to adjust their learning and monitoring to the task. This leads to the hypothesis that the delayed-keyword procedure will also be more effective for OSLs than for OPLs, resulting in a reduction in the difference between the media from the first studied texts. For the groups who studied from their non-preferred medium this study was exploratory. In Experiment 1, these groups did not benefit from practice. We were curious to see whether they would be able to benefit from the opportunity to access more-predictive cues.

3.1 Method

3.1.1 Participants. Seventy-six students (M_{age} = 26.8 years; 45% females) were drawn from the same population used for Experiment 1. The participants were randomly assigned to OSL (N = 37) and OPL (N = 39) groups.

3.1.2 Materials. The materials were those used for Experiment 1.

3.1.3 Procedure. The procedure resembled that of Experiment 1, except that (1) participants generated four keywords for each text; and (2) the keyword generation, predictions, and tests were delayed. The session started by demonstrating the entire procedure using the shorter text, including the keyword generation.

Participants read two texts consecutively with a time limit of seven minutes per text. Then the title of the first text was presented on the screen and the participants were instructed to write four keywords in designated spaces on the screen. Immediately afterward, participants filled in their POPs and took the test for the first text. Next, the same procedure (keywords, POPs, test) was performed for the second text. The entire procedure was then repeated for the third and fourth texts, and then for the fifth and sixth. Finally, the participants filled in the self-report questionnaire, as in Experiment 1.

3.2 Results

In the self-report, 60% of the participants stated that they would print the text for thorough learning. Regarding perceptions of study effectiveness, the percentages expecting better outcomes for paper, screen, or no difference were 62%, 3% ($N = 2$), and 35%, respectively (one participant did not answer this question). These preferences are consistent with those found in Experiment 1.

3.2.1 The effect of delayed keywords

We first analyzed the results for Pair 1, in order to examine whether the keywords procedure helped screen learners overcome screen inferiority even before practicing the task. A two-way Measure (POP vs. score) × Medium (screen vs. paper) ANCOVA yielded a main effect for the measure, $F(1, 72) = 14.82$, $p < .001$, $\eta_p^2 = .17$, where POPs were higher than test scores. There was no main effect for the medium, and also no interactive effect, meaning that we found equivalent test scores, POPs, and overconfidence in both media, with no screen inferiority.

Using delayed keyword generation provides an opportunity for in-depth processing, but also opens the door to forgetting, because of the passage of time and

interference of the intermediate task. To observe the effects of these contradictory influences, a three-way ANCOVA of Experiment (1 vs. 2) × Measure (POP vs. score) × Medium (screen vs. paper) was conducted. It yielded a main effect of measure, $F(1, 155) = 46.18$, $p < .001$, $\eta_p^2 = .23$, indicating overconfidence in both experiments and also a main effect of the experiment, $F(1, 155) = 13.09$, $p < .001$, $\eta_p^2 = .08$, with both POPs and test scores in Experiment 2 lower than in Experiment 1 regardless of the study medium, $t(161) = 2.13$, $p < .05$ for POPs and $t(161) = 2.85$, $p = .005$ for test scores. Thus, the participants reflected in their POPs the greater challenge imposed by delayed tests.

3.2.2 The effect of practicing the delayed keywords

Experiment 1 showed different effects of practice in the two media on scores, POPs, and calibration bias. Thus, it was important to examine whether media differences came about as a consequence of the practice.

Fig. 3 presents the results in the same way as Fig. 2. The same analysis reported for Pair 1 was performed for Pair 2 and Pair 3. For Pair 2 this ANCOVA yielded only a main effect of the measure, $F(1, 72) = 7.50$, $p < .01$, $\eta_p^2 = .09$, indicating overconfidence. For Pair 3 the similar analysis did not reveal the significant overconfidence found in Pair 1 and Pair 2, but only a nearly significant difference, $F(1, 72) = 2.76$, $p = .10$, $\eta_p^2 = .04$, indicating almost well-calibrated POPs.

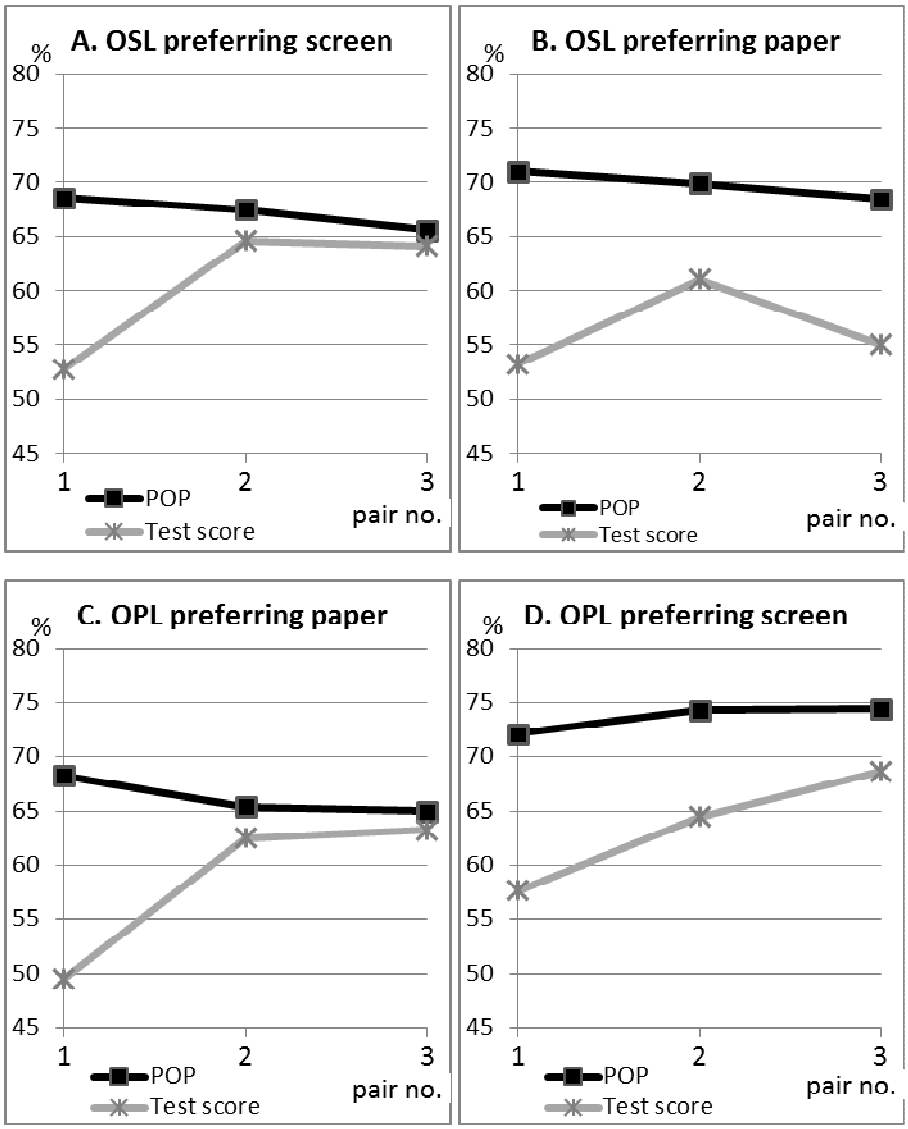

Fig. 3. Mean test scores and predictions of performance (POP) for the three text pairs by their study order, in division by study medium (On-Screen Learners – OSL, On-Paper Learners – OPL) and medium preference.

We turn now to the effect of medium preference. In particular, we examined whether groups who studied from their non-preferred medium would benefit from greater utilization of more-predictive cues. A two-way ANCOVA of Practice (pair 1, 2, 3) × Preferred Medium (screen vs. paper) on test scores for OSL yielded no effects, all Fs < 1, suggesting no improvement regardless of the medium preference. Interestingly, when SAT scores were removed from the model, a pronounced improvement with practice was

exposed, $F(2, 68) = 4.78$, $p = .01$, $\eta_p^2 = .12$. The ANCOVA for OPLs yielded only a main effect of order, $F(2, 35) = 4.66$, $p = .01$, $\eta_p^2 = .12$, suggesting that for them as well preference had no effect on improvement with practice. Thus, both groups improved their scores with practice, but for OSLs this improvement was dependent on participants' general cognitive ability, as reflected by SAT scores. Analyzing the effects of the medium on Pair 3 only for those who studied from their preferred medium eliminated screen inferiority altogether, $F < 1$, while for those who studied from their non-preferred medium, the screen inferiority that was eliminated in Pair 1 marginally emerged again, $F(1, 41) = 3.77$, $p = .06$, $\eta_p^2 = .08$. Thus, utilization of in-depth processing and practicing the task regenerated a medium effect arising from the differential ability of participants preferring different media to benefit from these aids.

A similar analysis was performed regarding POP. A two-way ANOVA of Practice (pair 1, 2, 3) × Preferred Medium (screen vs. paper) on POP for OSLs yielded, again, no significant effects, all Fs < 1.4. These findings suggest that the participants tended to stick to their initial predictions, even those who achieved higher scores after practice. The same ANOVA for OPLs yielded a main effect of preferred medium, $F(1, 37) = 4.42$, $p < .05$, $\eta_p^2 = .11$, meaning that OPLs who studied from their preferred medium were more modest in their POPs. The difference became significant for Pair 2, $t(37) = 2.13$, $p < .05$, and Pair 3, $t(37) = 2.32$, $p < .05$.

Turning to overconfidence, a three-way ANCOVA, similar to that performed in Experiment 1, of Practice (pair 1, 2, 3) × Medium (screen vs. paper) × Preferred Medium (screen vs. paper) on overconfidence yielded a nearly significant triple interaction, $F(2, 138) = 2.85$, $p = .06$, $\eta_p^2 = .04$. Overconfidence in Pair 1 was similar for all groups, but in

Pair 3 it was dependent on both the medium and medium preference. OSLs and OPLs who studied from their preferred medium (Panel A and Panel C) were perfectly calibrated at Pairs 2 and 3, $p > .4$ for the difference from zero. OPLs who preferred studying from screen (Panel D) were nearly well-calibrated at Pair 3, $p = .1$. OSLs preferring paper (Panel B) remained overconfident, $t(24) = 5.60$, $p < .005$, despite in-depth processing and practice with the task.

Finally, we examined whether participants who practiced the delayed keywords procedure (Experiment 2) could overcome the challenge imposed by the test delay and perform comparably to those who practiced the task with no delay (Experiment 1). A three-way ANCOVA of Experiment (1 vs. 2) × Medium (screen vs. paper) × Medium preference (screen vs. paper) over Pair 3 test scores yielded no main effect of the experiment, $F < 1$, and a significant interaction of medium and medium preference, $F(1, 150) = 4.22$, $p < .05$, $\eta_p^2 = .03$. Over both experiments, at Pair 3, OSLs studying from their preferred medium scored higher than those who preferred paper, $F(1, 70) = 6.36$, $p = .01$, $\eta_p^2 = .08$, while for OPLs, the medium preference showed no effect, $F < 1$. A similar ANCOVA over overconfidence produced similar findings. No main effects were found, Fs < 1, and again there was an interaction of medium and medium preference, $F(1, 150) = 8.40$, $p < .005$, $\eta_p^2 = .05$. For both OSLs and OPLs, those who studied from their preferred medium were less overconfident. For OSLs this effect was only nearly significant, $F(1, 70) = 3.45$, $p = .07$, $\eta_p^2 = .05$, while for OPLs it was significant $F(1, 83) = 4.43$, $p < .05$, $\eta_p^2 = .05$. Thus, over the two experiments, studying from one's preferred medium resulted in better calibration.

3.3 Discussion

Experiment 2 encouraged in-depth processing and thereby eliminated screen inferiority at the beginning, before participants practiced the task. The toll of this method was a reduction in test scores at Pair 1 when compared to Experiment 1. An important finding is participants' recognition of the difficulty of the delayed test, which was manifested in lower POPs. OPLs, regardless of their preferred study medium, utilized the practice cycles and achieved better test scores and attenuated overconfidence. For OSLs, in contrast, the improvement was only for those who preferred screen learning. Those who were forced to learn on-screen, but who would have preferred to study on paper if given the choice, did not benefit from practicing the in-depth processing method of keyword generation. Notably, relative to Experiment 1, in Experiment 2 more groups improved their scores and calibration by practicing the task and achieved similar results as in the immediate test in Experiment 1.

4. General Discussion

Several studies which examined the effects of the medium on continuous text learning have suggested that computerized learning generates contextual cues that impede cognitive and metacognitive processes (Ackerman & Goldsmith, 2011; Ackerman & Lauterman, 2012; Morineau et al., 2005). The present study provides evidence that the natural learning process tends to be shallower on screen than on paper, and offers two potential means of partially eliminating this barrier by guiding on-screen participants to learning processes which are more appropriate to the knowledge level required at test, and which seem to take place spontaneously when learning on paper. In addition, it

extends the research dealing with effects of the medium on learning by highlighting the importance of considering personal study preferences.

This study examined two methods aimed at reducing screen inferiority in performance and overconfidence relative to learning from paper. The first, practicing the task, attenuated the overconfidence of participants reading from their preferred medium in Experiment 1. This may hint that participants succeeded in utilizing more-predictive cues. Importantly, practicing the task did not enhance performance when the medium was not the preferred one. Using their non-preferred medium seems to prevent participants from recruiting the mental effort required to achieve cognitive and metacognitive processes as effective as those demonstrated by OSLs preferring screen and OPLs preferring paper.

Experiment 2 employed delayed keyword generation and testing, a method previously shown to improve performance and resolution (Thiede et al., 2003, 2005). By repeating the procedure six times, we could examine the immediate as well as accumulated benefits of this in-depth study procedure. Delayed keyword generation did eliminate the differences between OSLs and OPLs in both test scores and overconfidence already in Pair 1. Still, this method also created different patterns of learning which depended on participants' medium preference. Throughout, OPLs who studied from their preferred medium were more modest in their predictions than were OPLs who preferred studying on screen. After practicing the delayed keywords procedure, three out of four groups were well-calibrated. Screen inferiority remained despite in-depth processing and practice with the task only among those who studied on screen but preferred studying on paper.

Notably, after practicing the delayed keyword-generation procedure, participants overcame the challenge imposed by the delayed test and achieved similar test scores as those who practiced the immediate test. As delayed tests are more common than immediate ones, these findings suggest that combining practice with methods of in-depth processing can improve performance more effectively than either one alone.

A more general objective of this study was to examine whether the two methods – practice and delayed keyword generation – would improve calibration. Calibration is highly important if people are to manage their effort investment effectively. In particular, overconfidence is expected to lead people to think they have achieved an adequate level of knowledge and to cease effort investment too early, while their actual knowledge is still too low to satisfy their goals (Nelson & Narens, 1990; see Ackerman & Goldsmith, 2011, Fig. 1 for an illustration). Given that the metacognitive literature has dealt more with improving resolution than with improving calibration, the present study contributes to the literature by demonstrating that both methods are effective in attenuating overconfidence.

In both Experiment 1 and Experiment 2, overconfidence was attenuated with practice when the participants studied from their preferred medium. However, the results demonstrate two different mechanisms for calibration improvement. In the OSL group of Experiment 1 and in Experiment 2, regardless of the medium, improved learning regulation resulted in better performance for the last texts than the first ones, despite the constant time frame. This demonstrates improvement in study efficiency. Overconfidence was attenuated because POPs did not reflect this improvement, and even tended to fall. The remaining group, OPLs preferring paper in Experiment 1, did not improve their

learning regulation with practice, but adjusted their POPs better to their actual performance level. Both these directions for reducing calibration bias can be found in the literature, but mainly in studies involving memorization tasks (e.g., Koriat et al., 2002). The ability of practice to reduce learners' illusions of competence in memorization tasks was found to be list-dependent, meaning that it did not spontaneously transfer to new lists (e.g., Koriat & Bjork, 2006). The present study, with the more complex task of text learning, demonstrates transfer to new study items.

Koriat (1997) proposed that early in learning, subjects rely on less-reliable cues like the perceived intrinsic difficulty of items, but after repeated study of the same list they rely on more valid cues, such as encoding fluency. Future studies are called for to examine whether the improvement in calibration found here stems from utilization of more-reliable cues that allow better assessment of knowledge, or from processes that affect performance at test and its judgments independently. Considering these possibilities and understanding the processes that underlie calibration bias are generally important for understanding how to improve knowledge monitoring.

Beyond offering methods that may improve calibration, the present study highlights the factors of context and personal preference, which have not previously been taken into account in research on methods for improving metacognitive accuracy. We call for future studies on metacognitive accuracy to consider these factors when developing additional methods for improving it. Notably, in the present study, those who studied from their non-preferred medium benefited from practice (in terms of both increased knowledge and improved calibration) in the last texts only in limited conditions (among paper learners who preferred screen learning in Experiment 2). A better understanding of

the effects of context and personal preference on calibration may produce more effective ways to help those who struggle to learn on screen but not paper, or vice versa.

From a broader perspective, the literature dealing with self-regulated learning has highlighted other factors that affect learning effectiveness in computerized learning environments. With the present study, which offers methods for improving spontaneous assessments of knowledge, we hope both to improve the theoretical understanding in this area and to promote the development of practical methods for enhancing learning. For example, recognition of the possible interaction between contextual factors and individual differences, those considered here as well as others (e.g., Antonietti & Colombo, 2008; Colombo & Antonietti, 2011; Shaw & Marlow, 2000), may guide the development of computerized learning environments which can be adapted to suit learners' needs and preferences (see Vandewaetere, Desmet, & Clarebout, 2011, for a review). Another area involves computerized support for enhancing knowledge assessment (e.g., Roll, Aleven, McLaren, & Koedinger, 2011), where recognition of screen inferiority and ways to overcome it may promote the development of more effective support. Yet another direction for development relates to the combination of spontaneous and conscious metacognitive processes within discipline-specific learning research, such as science education (e.g., Zohar & Barzilai, in press). It is hoped that the current findings will lead to further development of discipline-specific methods for enhancing learning regulation inspired by the general methods we found effective here.

In sum, the consistent screen inferiority found in performance and overconfidence can be overcome by simple methods, such as practice and guidance on in-depth processing, even to the extent that some learners become able to perform as well on

screen as on paper. The findings have clear implications. First, software designers and policy makers in numerous contexts should take into account the differences between the media in the quality of learners' monitoring and regulation. Second, the principle of improved reliability of the cues used for monitoring, which guided us in choosing the methods tested in this paper, should be taken into account when designing computerized environments that involve extensive textual sections. In addition, the observed differences between the media in the effectiveness of such methods should draw attention to the fact that some methods reported in the literature were examined on only one medium, either screen or paper. From a theoretical perspective, the found effects draw attention to the effects of the context and learner preference on learning regulation and outcomes, beyond the factors, such as students' learning skills and attributes of the study task, that are traditionally examined.

References

Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. *Journal of Experimental Psychology: Applied*, *17*(1), 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*(5), 1816–1828.

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.*

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

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

Antonietti, A., & Colombo, B. (2008). The effects of computer-supported learning tools: A bi-circular bi-directional framework. *New Ideas in Psychology*, 26, 120–142.

Antonietti, A., Colombo, B., & Lozotsev, J. (2008). Undergraduates' metacognitive knowledge about the psychological effects of different kinds of computer-supported instructional tools. *Computers in Human Behavior*, *24*, 2172–2198.

Azevedo, R. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational Psychologist, 40*, 199–209.

Baker, J. M. C., & Dunlosky, J. (2006). Does momentary accessibility influence metacomprehension judgments? The influence of study – judgment lags on accessibility effects. *Psychonomic Bulletin & Review, 13*(1), 60–65.

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

Buzzetto-More, N., Sweat-Guy, R., & Elobaid, M. (2007). Reading in a digital age: E-books: Are students ready for this learning object? *Interdisciplinary Journal of Knowledge and Learning Objects*, *3*, 239–250.

Colombo, B., & Antonietti, A. (2011). Self-regulated strategies and cognitive styles in multimedia learning. In G. Dettori & D. Persico (Eds.), *Fostering self-regulated learning through ICT* (pp. 54–70). New York: Information Science Reference.

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

Dunlosky, J., & Rawson, K. A. (2005). Why does rereading improve metacomprehension accuracy? Evaluating the levels-of-disruption hypothesis for the rereading effect. *Discourse Processes, 40*, 37–55.

Dunlosky, J., & Tauber, S. K. (in press). Understanding people's metacognitive judgments: An isomechanism framework and its implications for applied and theoretical research. In T. Perfect & S. Lindsay (Eds.), *Handbook of Applied Memory*. Thousand Oaks, CA: Sage.

Fukaya, T. (in press). Explanation generation, not explanation expectancy, improves metacomprehension accuracy. *Metacognition and Learning*.

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition,* 15, 84-93.

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*, 334–372.

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., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*(1), 160-170.

Jamali, H. R., Nicholas, D., & Rowlands, I. (2009). Scholarly e-books: the views of 16,000 academics: Results from the JISC national e-book observatory. *Aslib Proceedings, 61(1),* 33-47.

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

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

Koriat, A., & Bjork, R. A. (2006). Mending metacognitive illusions: A comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1133-1145.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*(2), 147-162.

Lauterman, T., & Ackerman, R. (2013). Overcoming the screen inferiority in text learning. In M. Knauff, M., Pauen, N., Sebanz, & I. Wachsmuth (Eds.) *Proceedings of the 35th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society (pp. 2838-2842).

Liu, Z. (2005). Reading behavior in the digital environment: Changes in reading behavior over the past ten years. *Journal of Documentation, 61*(6), 700-712.

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.*

Maki, R. H., & Serra, M. (1992). Role of practice tests in the accuracy of test predictions on text material. *Journal of Educational Psychology, 84, 200-210*.

Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research*, *58*, 61-68.

Metcalfe, J. (1998). Cognitive optimism: Self-deception or memory-based processing heuristics? *Personality and Social Psychology Review, 2*(2), 100-110.

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

Morineau, T., Blanche, C., Tobin, L., & Guéguen, N. (2005). The emergence of the contextual role of the e-book in cognitive processes through an ecological and functional analysis. *International Journal of Human–Computer Studies, 62,* 329–348.

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.

Rawson, K. A., & Dunlosky, J. (2002). Are performance predictions for text based on ease of processing? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 69–80.

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2007). Designing for metacognition—applying cognitive tutor principles to the tutoring of help seeking. *Metacognition and Learning*, *2*, 125–140.

Shaw, G. P., & Marlow, N. (2000). The role of student learning style, gender, attitudes and perceptions on information and communication technology assisted learning. *Computers and Education, 33*, 223–234.

Spencer, C. (2006). Research on learners' preferences for reading from a printed text or from a computer screen. *Journal of Distance Education*, *21*, 33–50.

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, and Cognition, 31*, 1267–1280.

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J.S. (2009). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of Metacognition and Self-Regulated Learning* (pp. 85–106). New York: Routledge.

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

Vandewaetere, M., Desmet, P., & Clarebout, G. (2011). The contribution of learner characteristics in the development of computer-based adaptive learning environments. *Computers in Human Behavior*, *27*, 118–130.

Veenman, M. V. J., Prins, F. J., & Elshout, J. J. (2002). Initial learning in a complex computer simulated environment: The role of metacognitive skills and intellectual ability. *Computers in Human Behavior, 18*, 327–342.

Weaver, C. A., III, & Bryant, D. S. (1995). Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text. *Memory & Cognition, 23*, 12–22.

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: Lawrence Erlbaum Associates, Inc.

Woody, W. D., Daniel, D. B., & Baker, C. A. (2010). E-books or textbooks: Students prefer textbooks. *Computers & Education, 55*(3), 945-948.

Zohar, A., & Barzilai, S. (in press). A review of research on metacognition in science education: Current and future directions. *Studies in Science Education*, 1–49.