October 2017

Sidi, Y, Shpigelman, M., Zalmanov, H., & Ackerman, R. (2017). Understanding metacognitive inferiority on screen by exposing cues for depth of processing. *Learning and Instruction*, *51,* 61-73.

https://www.journals.elsevier.com/learning-and-instruction/

Understanding Metacognitive Inferiority on Screen by Exposing Cues for Depth of Processing

Yael Sidi, Maya Shpigelman, Hagar Zalmanov, and Rakefet Ackerman

Faculty of Industrial Engineering and Management,

Technion–Israel Institute of Technology, Haifa, Israel

Author note

Corresponding Author - E-mail: yaelsidi@gmail.com

The study was supported by a grant from the Israel Science Foundation (Grant No. 957/13) and by the General Research Fund at the Technion. We thank Tirza Lauterman and Tova Michalsky for insightful comments regarding earlier versions of this paper, and Meira Ben-Gad for editorial assistance.

Highlights

- Text-learning studies often find screen inferiority in knowledge and monitoring
- Minimizing the reading burden, we used brief but challenging problem-solving tasks
- Time pressure and framing the task as preliminary still yielded screen inferiority
- Metacognitive processes are sensitive to hints at the expected processing depth
- Eliminating screen inferiority is possible by cues calling for in-depth processing

Abstract

Paper-and-pencil learning and testing are gradually shifting to computerized environments. Cognitive and metacognitive researchers find screen inferiority compared to paper in effort regulation, test performance, and extent of overconfidence, in some cases, with unknown differentiating factors. Notably, these studies used reading comprehension tasks involving lengthy texts, which confound technology-related and cognitive factors. We hypothesized that the medium provides a contextual cue which leads to shallower processing on screen regardless of text length, particularly when task characteristics hint that shallow processing is legitimate. To test this hypothesis, we used briefly phrased yet challenging problems for solving on screen or on paper. In Experiment 1, the time frame for solving the problems was manipulated. As with lengthy texts, only time pressure resulted in screen inferiority. In Experiment 2, under a loose time frame, the same problems were now framed as a preliminary task performed before a main problem-solving task. Only the initial task, with reduced perceived importance, revealed screen inferiority similarly to time pressure. In Experiment 3, we replicated Experiment 1's time frame manipulation, using a problem-solving task which involved reading only three isolated words. Screen inferiority in overconfidence was found again only under time pressure. The results suggest that metacognitive processes are sensitive to contextual cues that hint at the expected depth of processing, regardless of the reading burden involved.

Keywords: Metacognition; Monitoring and control; Human-computer interaction; Problem solving; Effort regulation; Depth of processing

1. Introduction

Over recent decades, paper-and-pencil work has been shifting to computerized environments for many types of cognitive tasks in everyday contexts, including learning (e.g., MOOCs), work-related and academic screening (e.g., the GMAT and SAT), and surveys, as well as scientific research. This shift has been driven mainly by practical considerations, such as lower costs, automatic grading, and easy access to a wide audience, although, of course, computerized environments also allow novel task designs (e.g., Buhrmester, Kwang, & Gosling, 2011; Csapó, Ainley, Bennett, Latour, & Law, 2012; Dennis, Abaci, Morrone, Plaskoff, & McNamara, 2016; Mason & Suri, 2012; Quellmalz & Pellegrino, 2009).

While there is no doubt about the important advantages of computerized environments, the technological revolution compels us to ask what effects the medium might have on cognitive performance. Research in this area has yielded inconclusive results. On the one hand, there is evidence for both a subjective preference for paper (e.g., Holzinger et al., 2011; Kazanci, 2015; Mizrachi, 2015; Singer & Alexander, 2017; van Horne, Russell, & Schuh, 2016; Woody, Daniel, & Baker, 2010) and actual better performance on paper, relative to working on screen (e.g., Ben-Yehudah & Eshet-Alkalai, 2014; Daniel & Woody, 2013; Lin, Wang, & Kang, 2015; Mangen, Walgermo, & Brønnick, 2013). On the other hand, some studies have found no performance differences between the two environments, and several even point to screen superiority (e.g., Ball & Hourcade, 2011; Dennis et al., 2016; Holzinger et al., 2011; Margolin, Driscoll, Toland, & Kegler, 2013; Murray & Pérez, 2011; Salmerón & García, 2012). Finally, there are studies

which point to a discrepancy between learners' preference for digital environments and the actual learning outcomes (e.g., Singer & Alexander, 2017).

The inconsistency in the literature highlights the need for a thorough investigation of the conditions under which computerized learning should be expected to harm performance and those that allow eliminating this harmful effect. Our goal in the present study is to shed new light on conditions that lead to lower performance on screen than on paper and those that allow eliminating it, under the same technological conditions. To accomplish this, we used briefly phrased problem solving tasks and compared the results to the pattern of results found with tasks involving comprehension of lengthy texts, thereby generalizing and extending previous research.

In the following sections we delineate three types of explanations for the mixed results. We begin by weighing technological factors versus metacognitive regulation of mental effort. In particular, we elaborate on cues that legitimate shallow rather than in-depth processing in reading comprehension and problem solving. We then consider cognitive load as yet another factor that may contribute to the mixed results. Finally, we outline our study.

1.1. Technological versus regulatory explanations for screen inferiority

Lower performance on screen, when found, has been often explained in terms of technological disadvantages associated with electronic devices, such as screen glare, visual fatigue, and less-convenient navigation along the text relative to parallel task performance on paper (e.g., Benedetto, Drai-Zerbib, Pedrotti, Tissier, & Baccino, 2013; Moustafa, 2016; see Leeson, 2006, for a review). However, empirical evidence has been accumulating to suggest that this explanation is insufficient. First, such lower performance has been found

even with the latest e-books and tablets, which are presumed to overcome these technological limitations (e.g., Antón, Camarero, & Rodríguez, 2013; Daniel & Woody, 2013; Lin et al., 2015; see Gu, Wu, & Xu, 2015, for a review). Also pointing in the same direction is the perseverance of a paper preference even among experienced computers' users and young adults (e.g., Baron, 2013; Holzinger et al., 2011; Kazanci, 2015; Kretzschmar et al., 2013; Mizrachi, 2015). Finally, in several studies, lower performance on screen was found in some conditions but not in others (e.g., a pressured vs. loose time frame to complete a task), despite use of the same task on both media and comparable samples (Ackerman & Goldsmith, 2011; Ackerman & Lauterman, 2012; Lauterman & Ackerman, 2014). Technological disadvantages associated with screens should have taken their effect regardless of the condition. These findings hint that the main source for the found lower performance on screen may be cognitive in nature, rather than technology-related.

A potential cognitive explanation that has been gaining empirical support is based on differences in depth of processing between the media. For example, Daniel and Woody (2013) compared reading comprehension in e-textbooks and paper textbooks. While they found no medium effect on test scores, participants in the electronic conditions demonstrated less efficient work—they had to invest more time to achieve similar performance levels. Morineau, Blanche, Tobin, and Guéguen (2005) examined e-books and paper books as contextual cues for retrieval of learned information. They found that the mere presence of the e-book interfered with recall, while the presence of the paper book facilitated it. In addition, users' reports on their experience interacting with computerized environments convey a qualitatively different reading process on computer screens than on

paper, involving more interrupted work, attentional shifts, and multitasking, resulting in less time devoted to in-depth reading (Daniel & Woody, 2013; Hillesund, 2010; Liu, 2005). More recently, Mueller and Oppenheimer (2014) compared note taking using a laptop and regular handwriting. They found across three studies that participants who worked on screen used more verbatim note taking, compared to participants who worked on paper, even when participants were instructed not to take verbatim notes. This led to lower success rates for the screen group on recall and conceptual application questions. The authors suggested that working on laptops yielded shallower processing than writing on papers.

This explanation has recently received further support from studies dealing with self-regulated learning. These regulatory processes take place in parallel to the core cognitive processing during the performance of any cognitive task (e.g., storing information in memory during learning, interpreting a road sign during navigation, etc.). The metacognitive framework suggested by Nelson and Narens (1990) emphasizes in particular the central role of reliable monitoring in effective effort regulation. That is, knowledge monitoring guides spontaneous decisions regarding chosen learning strategies and allocation of time to the task. Unreliable monitoring is expected to yield ineffective regulatory decisions. For instance, overconfidence may mislead a learner to think prematurely that her study goal has been achieved and that no further activity is required (see Bjork, Kornell, & Dunlosky, 2013; Winne & Baker, 2013, for reviews). The present study employs a metacognitive framework, with the aim of illuminating conditions under which cognitive and metacognitive processes differ between the two media.

1.2. Media effects on Meta-Comprehension

Meta-comprehension is the research domain dealing with metacognitive aspects of reading comprehension tasks. In a series of meta-comprehension studies, Ackerman and colleagues found screen inferiority in three measures: the calibration of metacognitive monitoring in the direction of overconfidence; less effective effort regulation; and lower test scores (Ackerman & Goldsmith, 2011; Ackerman & Lauterman, 2012; Lauterman & Ackerman, 2014). Notably, in all these studies there were also conditions in which screen inferiority was not found. For instance, Ackerman and Goldsmith (2011) investigated the effect of time frame on working on screen versus on paper. No significant difference between the media was found under a limited time frame with a sample from a population with a strong paper preference. However, when the participants were free to regulate their learning by themselves, those who studied on screen showed overconfidence and did not benefit from the extra time they invested, while those who studied on paper improved both their monitoring calibration and test scores.

Ackerman and Lauterman (2012) replicated this study with a sample of technology-savvy students, characterized by an attenuated paper preference. They found highly similar screen inferiority, but only under time pressure. Notably, screen inferiority was found only when the time limit was known in advance, but not when participants were interrupted unexpectedly after the same amount of study time. Time pressure has been associated in the literature with compromising on one's goal (Thiede & Dunlosky, 1999). This notion leads us to appreciate the adjustment made by paper participants but not by screen participants. Specifically, participants who worked on paper improved their learning efficiency without compromising on their goals when the task characteristics called for it, presumably by

recruiting extra mental effort. Conversely, participants working on screen had similar efficiency with and without time pressure, even though the time frame was known in advance.

Lauterman and Ackerman (2014) replicated the screen inferiority found by Ackerman and Lauterman (2012) under time pressure. Subsequently, they demonstrated two readily applicable methods for overcoming screen inferiority, gaining experience with the challenging learning task and a requirement to generate keywords summarizing the essence of the text after a delay (adapted from Thiede, Anderson, & Therriault, 2003). The findings of this study also suggest that the default processing on screen under time pressure is shallower than on paper, as an external trigger was required to eliminate screen inferiority. Importantly, this research suggests that employing simple task characteristics allow eliminating screen inferiority altogether.

The studies mentioned above examined effects of the medium on cognitive performance by using reading comprehension tasks, involving texts spread over a whole page or even several pages (e.g., 1000-1200 words in Ackerman & Lauterman, 2012; 858 word in Ben-Yehudah & Eshet-Alkalai, 2014; 1400-1600 words in Mangen et al., 2013). However, the lengthier the text, the more it is susceptible to the technological disadvantages associated with screen reading (e.g., eye strain). Thus, these studies confound technological disadvantages and in-depth processing.

In the present study we addressed this confound by reducing dramatically the room for technological factors to take effect, without scaling down the cognitive effort required by the task, by using briefly phrased yet challenging problem solving tasks. In order to delve

into the metacognitive processes involved, we employed the meta-reasoning framework (Ackerman & Thompson, 2015).

1.3. Media effects on Meta-Reasoning

Meta-Reasoning is an emerging domain applying the metacognitive framework to problem solving, by examining judgments and regulatory decisions that accompany performing reasoning challenges (see Ackerman & Thompson, 2015, for a review). Overall, the general finding in meta-reasoning studies is that problem solvers tend to be overconfident (Ackerman & Zalmanov, 2012; Prowse Turner & Thompson, 2009; Shynkaruk & Thompson, 2006). Just as in learning, overconfidence may lead people to conclude prematurely that they have found a satisfactory solution to the problem and halt their solving efforts (Ackerman, 2014; Evans, 2006). Given the increasing use of computerized screening exams and other high-stakes problem-solving contexts, exposing factors that affect metacognitive processes is important for practical considerations. However, it also has theoretical importance, as within the meta-reasoning literature most studies consider cues that are inherent to the task itself (e.g., familiarity of question terms; Reder & Ritter, 1992), its performance (e.g., answer fluency—the speed with which the answer is produced; Thompson et al., 2013), or individual differences (e.g., math anxiety, Morsanyi, Busdraghi, & Primi, 2014). Interactions with external conditions, such as media, are rarely considered.

Recently, Meyer et al. (2015) reviewed a collection of studies which compared brief problem solving tasks presented in regular fonts or in hard to read fonts (e.g., easy to read vs. *hard to read*). The font manipulation was meant to increase depth of processing (see Thompson et al., 2013), although it was recently found that in most cases it does not affect

performance (Meyer et al., 2015; see Kühl & Eitel, 2016, for a review). The reviewed studies were conducted either on screen or on paper. Meyer et al. examined the media as a secondary factor in their review and concluded that the media did not make a difference and did not interact with font legibility. Similarly, no global media effect on problem solving was found by Sidi, Ophir, and Ackerman (2016) with the same brief task, which takes 1-2 minutes to perform. Notably, in addition to the font legibility manipulation, this study had the media as a manipulated factor and included confidence ratings in one of the experiments. When measuring confidence, Sidi et al. found that font legibility affected performance on both media: Performance was improved on screen by the hard to read fonts, while on paper the opposite effect was found. Importantly, on screen, confidence ratings were not sensitive to performance differences between the regular and less-legible fonts, while on paper they reliably reflected the performance difference between the presentation fonts. This finding generalizes the finding of less reliable metacognitive monitoring on screen compared to paper, even in this brief task, as previously found with lengthy texts. In the present study we aimed to examine the generalizability of this particular insensitivity of confidence ratings to performance differences on screen, and shed more light on the effects of cues for depth of processing on screen and on paper.

1.4. Cognitive load

Considering problem solving tasks and working under time pressure brings to the fore the Cognitive Load Theory (Sweller, 1976), which was not taken into account in the previous studies examining media effects on effort regulation. This theory has been very influential in providing guidelines for instructional design for developing problem solving skills in educational contexts (see Schnotz & Kürschner, 2007, for a review). In particular, it

has been considered in light of recent computerized learning environments which incorporate elements such as hypertexts and animation within study materials. Notably, the results are mixed. Höffler and Leutner (2007) found in a meta-analysis a medium-sized overall advantage of instructional animations over static pictures which was explained in terms of reducing cognitive load. However, they also found several moderators focusing the found advantage to representational animations, highly realistic animations, and/or when procedural-motor knowledge is to be acquired. In line with these findings, other studies suggested that technology-based features may overload the cognitive system if not employed carefully (e.g., DeStefano & LeFevre, 2007; Hollender, Hofmann, Deneke, & Schmitz, 2010). For example, animations can potentially increase cognitive load by distracting the learner from essential information, or due to their transient nature, which requires the learner to store more information in working memory (Ayres & Pass, 2007). In light of the findings of media effects on reading comprehension, without any technology-based features, the present study goes a step back, and considers the option that the mere presentation media is an interfering factor, generating extraneous load, even in tasks that can be presented in the same way on screen and on paper.

Cognitive load considerations are particularly relevant for analyzing work under time pressure. On the one hand, time pressure has been strongly associated with an increase in extraneous cognitive load and a reduction in performance (Barrouillet, Bernardin, Portrat, Vergauwe, & Camos, 2007; Paas & Van Merriënboer, 1994). On the other hand, there were also findings of unharmed performance, even under severe time pressure, suggesting on "good" cognitive load (germane load, Sweller, Van Merriënboer, & Paas, 1998). For instance, Gerjets and Scheiter (2003, study 4) examined the effect of time pressure during the

learning stage of a problem solving task using multiple instructional conditions. Based on the Cognitive Load Theory, they expected participants under time pressure to skip some of the instructional material, resulting in lower performance. However, time pressure did not impair learning in their study. The authors suggested that time pressure can increase germane load, guiding people to make effective strategic adjustments. This explanation resembles the metacognitive explanation reviewed above for adjustment to time pressure, which was found only for paper, but not for screen (Ackerman & Lauterman, 2012; Lauterman & Ackerman, 2014). Thus, a similar inferential effect can be made for cognitive load: We suggest that the media may interact with the effects of time pressure on cognitive load, an idea that as far as we know was not considered before. Notably, discussions of metacognition in the context of cognitive load are mostly related to explicit reflection on study strategies (e.g. Valcke, 2002), which is out of the scope of the present study.

In the present study, we employed a time frame manipulation with problem solving similarly to that examined before with reading comprehension, as described above. However, we also employed another manipulation, perceived importance of the task, to examine whether screen inferiority is associated to an increased cognitive load which occurs under time pressure, or can be found in other contexts as well.

1.5. Overview of the present study

To minimize the role of technological factors, in Experiment 1 we replicated the time frame procedure used before with lengthy texts (Ackerman & Lauterman, 2012; Lauterman & Ackerman, 2014), but here we used challenging problem solving tasks, which were briefly phrased. For differentiating between the cognitive load and the regulatory explanation, in Experiment 2 we manipulated perceived importance of the task. Lower perceived importance

was expected to serve as another cue for shallow processing that does not involve an increase in cognitive load. In Experiment 3, we used again the time frame procedure, but with even shorter problems involving reading only isolated words, for eliminating the reading comprehension component from the task. We hypothesized that computerized environments lead people to adapt shallower processing than paper environments under manipulations that legitimate compromise, regardless of the reading burden or the cognitive load generated by time pressure. Thus, we expected screen inferiority to be found in all cases involving cues that legitimate shallower processing, in line with the regulatory explanation.

2. Experiment 1

In order to examine our hypotheses, we chose extremely challenging logic problems which are brief in terms of their reading burden (see details in the Materials section) which we adapted from Ackerman, Leiser, and Shpigelman (2013). The text of each problem included less than 100 words—far less than the several pages used in the reading comprehension studies mentioned above. We examined the effect of the medium (screen or paper) and time frame (pressured versus loose) on response time, confidence, overconfidence, problem-solving efficiency (correct solutions per hour), and ultimate success rate. Using a similarly technologically-savvy population, we predicted that Ackerman and Lauterman's findings of screen inferiority under time pressure and media equivalence under a loose time frame would be replicated, despite the substantially different task.

As taking the time frame into account during the task presents a burden in itself both groups worked under predefined time frames. The time allotted for the loose time frame (LTF) group was defined based on a previous study that used the same problems with a

sample from the same population (Ackerman et al., 2013). The time frame for the time pressure (TP) group was 66% of the time allowed for the LTF group. The instructions presented the time frame as loose or pressured, accordingly. Participants were required to complete the entire task in the allotted time.

2.1. Method

2.1.1. Participants

One hundred and three undergraduate engineering students at the Technion–Israel Institute of Technology were randomly assigned to work on screen or on paper, under time pressure or a loose time frame (N = 22-31 per group; M_{age} = 24.4, SD = 2.4; 39% females). Participants reported not having any learning disabilities. Notably, this sample—which was drawn from the same population as in Ackerman and Lauterman (2012)—is highly familiar with computerized environments[1] and has high cognitive ability (the Technion's undergraduate programs typically require SAT scores in the top 20%).

2.1.2. Materials

The materials were six logic problems used with a sample from the same population by Ackerman et al. (2013, Experiment 2). See Appendix. These problems were designed to be highly challenging for the target population, with success rates lower than 20%. The problems consisted of 77 Hebrew words on average.

[1] A self-report survey on this population (N = 247), which was conducted parallel to the present study, revealed that computerized environments are an integral part of the students' daily life. In particular, participants were accustomed to using computers from childhood (M = 8.9 years old, SD = 3) and reported currently using them for a large portion of each day (M = 5.79 hours, SD = 2.7).

2.1.3. Procedure

The experiment was administered in groups of up to eight participants in a small computer lab. All participants in each experimental session worked under the same condition and faced all six problems successively, randomly ordered for each participant. For the screen group, when the "Start solving" button was pressed, the problem appeared on the screen, with an empty space below for entering the solution. Participants had at their station blank sheets of paper and pens for scribbling or sketching while solving. Pressing the "Continue" button exposed a confidence rating scale (0-100%). Participants indicated their confidence rating by dragging an arrow along the scale. Then they indicated whether they knew the problem in advance (yes/no) and clicked "Next" to move on to the next problem. Response time was measured from when participants clicked "Start solving" to when they clicked the "Next" button.

For the paper group, each problem appeared on a separate page. The pages included space for scribbling and for writing the answer. A horizontal scale (0-100%) for confidence ratings appeared at the bottom of the page, similarly to its appearance on the screen, along with the yes/no "advance knowledge" question. The participants indicated their confidence rating by marking a vertical line on the scale. The pages were prearranged in a pile for each participant, upside down; participants turned over one page at a time, and turned each page over into a second pile when they completed it. So that we could measure response time, participants clicked "Start solving" and "Next" buttons on an otherwise empty screen at the start and end of each problem. This was their only interaction with the computer.

In all groups, the participants moved from one problem to the next without returning to previous ones. The TP participants had 24 minutes to solve the entire problem set, with time

reminders at 5, 20, and 23 minutes. The experimenter emphasized that this time frame was pressured and that it allowed about 4 minutes per problem. Participants were explicitly instructed to manage their time to allow solving the entire problem set. For the LTF group, the time frame was 36 minutes, with reminders at 7, 30, and 35 minutes. The experimenter explained that this time frame allowed relaxed work, but that participants should pay attention to the time and ensure they completed the entire problem set.

2.2. Results and discussion

Table 1 summarizes the medium comparisons. It provides a bird's eye view over all the measures across the three experiments. The means and analyses' results appear in the results report of each experiment, in the figures or in the detailed description.

Table 1. Summary of medium comparisons (S – Screen, P – Paper) in the three experiments.

Experiment and condition	Response time	Confidence	Overconfidence (lower is better)	Efficiency	Success rate
Experiment 1 – Challenging logic problems					
Time pressure	S ≈ P	S ≈ P	S > P	S < P	S < P
Loose time frame	S ≈ P	S ≈ P	S ≈ P	S > P	S > P
Experiment 2 – Metacognitive Transfer Paradigm (MTP)					
Initial problems	S < P	S ≈ P	S > P	S ≈ P	S < P
Transfer problems	S < P	S ≈ P	S ≈ P	S ≥ P	S ≈ P
Experiment 3 – Compound Remote Associate (CRA) problems					
Time pressure	S ≈ P	S ≈ P	S > P	S ≈ P	S ≈ P
Loose time frame	S ≈ P	S ≈ P	S ≈ P	S ≈ P	S ≈ P

< or > A statistically significant difference, $p < .05$,
≥ A marginal difference, $p = .053$
≈ A non-significant statistical difference

Overall the data from Experiment 1, eighteen problems (3%, one per participant) were known in advance. These problems were removed from the analyses. We started our analyses by testing for a medium effect on sketching while solving the problems, as a potential indicator for depth of processing. Sketching was measured as dichotomy—any type of scribble was considered a sketch and was counted as "yes", while solving without any scribble was counted as "no". Analysis of Variance (ANOVA) for effect of the Medium (screen vs. paper) and the Time Frame (TP vs. LTF) on the number of solutions for which sketches were used revealed no difference between the media, $F < 1$. Participants scribbled or sketched to a lesser extent under TP ($M = 28\%$, $SD = 34$) than under LTF ($M = 42\%$, $SD = 34$), $F(1, 99) = 8.93$, $MSE = 0.52$, $p = .004$, $\eta_p^2 = .083$, but there was no interaction with the medium, $F < 1$. Thus, working on screen clearly did not lead participants to avoid sketching as a problem-solving aid. A similar ANOVA on response times revealed only the obvious difference between the time frames, with shorter response times under TP ($M = 2.3$ min., $SD = 0.3$) than under LTF ($M = 3.3$ min., $SD = 0.5$), $F(1, 99) = 153.55$, $MSE = 0.47$, $p < .001$, $\eta_p^2 = .608$. Thus, participants took the opportunity to sketch out their solution ideas to a similar extent regardless of the medium, with no difference in the time they invested in each problem.

Success was scored as correct or incomplete/wrong. As intended, the problems were highly challenging, resulting in low success rates ($M = 20.7\%$, $SD = 15.2$); see Figure 1. Nevertheless, as expected, the pattern of success rates when comparing the four groups was highly similar to that found by Ackerman and Lauterman (2012) with a reading comprehension task. An ANOVA as above on success rates revealed no effect of the medium, $F < 1$, an effect of the time frame, $F(1, 99) = 8.86$, $MSE = 188.45$, $p = .004$, $\eta_p^2 =$

.082, and an interaction effect, $F(1, 99) = 14.69$, $MSE = 188.45$, $p < .001$, $\eta_p^2 = .129$. Comparing the time frames on screen revealed lower success rates under TP ($M = 11.3\%$, $SD = 11.7$) than under LTF ($M = 29.9\%$, $SD = 13.9$), $t(53) = 5.38$, $p < .001$, Cohen's $d = 1.38$. On paper, in contrast, the time pressure did not result in a compromise on performance, $t < 1$. Comparing the media within each time frame revealed lower success rates on screen ($M = 11.3\%$, $SD = 11.7$) than on paper ($M = 23.5\%$, $SD = 16.8$) under TP, $t(51) = 3.12$, $p = .003$, $d = 0.89$, while the opposite pattern was found under LTF, $t(48) = 2.29$, $p = .026$, $d = 0.66$, with higher success rates on screen ($M = 29.9\%$, $SD = 13.9$) than on paper ($M = 21.1\%$, $SD = 12.9$). The findings suggest that effective problem solving on screen is certainly possible, and results on screen can even be better than on paper when ample time is provided. However, time pressure reduced the success rate on screen but not on paper.

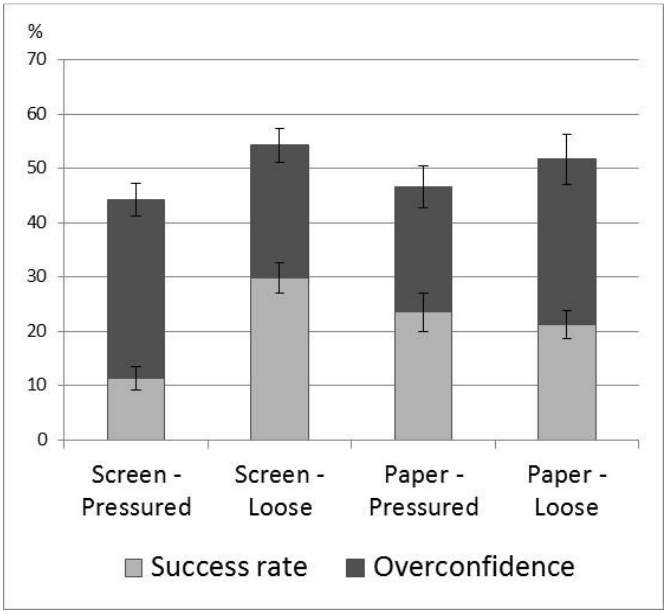

Figure 1. Success rates and overconfidence in solutions in Experiment 1. Confidence is represented by the top of the overconfidence bars. Error bars represent standard errors of the mean for the bar below them.

Similarly to findings with reading comprehension, confidence ratings did not necessarily correspond to the differences in success rates. An ANOVA on confidence revealed only a main effect of the time frame, $F(1, 99) = 5.70$, $MSE = 252.20$, $p = .019$, $\eta_p^2 = .054$, reflecting lower confidence under TP compared to LTF. There was no main effect for medium or an interaction effect (F's < 1). Overconfidence was calculated by comparing mean confidence ratings to mean success rates for each participant across the entire task. All groups showed a large degree of overconfidence, all ps < .001. To clarify the differential effects of the medium and time frame on the correspondence between confidence and success rates, we conducted an ANOVA on overconfidence. This analysis revealed no main effects for either the medium or time frame, both Fs < 1, but an interaction effect, $F(1, 99) = 4.60$, $MSE = 350.58$, $p = .034$, $\eta_p^2 = .044$. Comparing the time frames within each medium revealed a marginal difference on screen, $t(53) = 1.96$, $p = .054$, $d = 0.55$, with a tendency for greater overconfidence under TP ($M = 33.0$, $SD = 16.4$) compared to LTF ($M = 24.4$, $SD = 15.4$), while on paper, the time frame groups did not differ, $t = 1.19$, $p = .240$, $d = 0.35$. Comparing the media within each time frame revealed that under TP, overconfidence was greater on screen ($M = 33$, $SD = 16.4$) than on paper ($M = 23.1$, $SD = 18.4$), $t(48) = 2.04$, $p = .046$, $d = 0.58$, while under LTF there was no such difference, $t = 1$. Thus, the pattern of screen inferiority under time pressure but not under a loose time frame found before with reading comprehension tasks (Ackerman & Lauterman, 2012; Lauterman & Ackerman, 2014) was replicated here with briefly phrased but challenging logic problems.

In order to examine time management as a metacognitive control strategy, efficiency was calculated as the number of correct solutions per hour. An ANOVA on efficiency revealed no statistically significant main effects, F's < 1, but an interaction effect, $F(1, 99) =$

17.05, $MSE = 3.57$, $p < .001$, $\eta_p^2 = .147$. On screen, participants were less efficient under TP ($M = 1.7$, $SD = 1.8$) than under LTF ($M = 3.3$, $SD = 1.6$), $t(53) = 3.43$, $p = .001$, $d = 0.95$. On paper, in contrast, the work was more efficient under TP ($M = 3.8$, $SD = 2.7$) than under LTF ($M = 2.3$, $SD = 1.3$), $t(46) = 2.50$, $p = .016$, $d = 0.74$. A comparison between the media within each time frame revealed that under TP, work on screen was less efficient ($M = 1.7$, $SD = 1.8$) than on paper ($M = 3.8$, $SD = 2.7$), $t(51) = 3.33$, $p = .002$, $d = 0.95$, while under LTF, the pattern was reversed ,with more efficient work on screen ($M = 3.3$, $SD = 1.6$) than on paper ($M = 2.7$, $SD = 1.3$), $t(48) = 2.15$, $p = .015$, $d = 0.73$. Thus, high efficiency on screen can be achieved, but time pressure hinders it. On paper, in contrast, time pressure encourages efficient work.

To summarize, Experiment 1 replicated with a briefly phrased problem-solving task the findings of Ackerman and Lauterman (2012) with lengthy reading comprehension tasks. This outcome provides further evidence that although cognitive processing can be effective on screen, and sometimes even better than it is on paper, time pressure impedes cognitive processing on screen in particular. As for metacognitive monitoring, the insufficient adjustment of confidence ratings to performance variations, found only on screen, suggests that monitoring on screen was less tuned to factors that affect performance and to performance itself. Importantly, overconfidence was most apparent for screen solvers under time pressure. These findings generalize the findings of Sidi et al. (2016), in which the comparison within each medium was between fluent and disfluent fonts. Finally, with respect to effort regulation, while in a reading comprehension task (Ackerman & Lauterman, 2012) efficiency on screen remained constant in both time frames, here time pressure

reduced efficiency on screen. The high efficiency on paper in the time pressure condition highlights the faulty regulation of effort on screen.

3. Experiment 2

In Experiment 1, as predicted, time pressure resulted in screen inferiority, that is inferior monitoring, efficiency, and success rates on screen compared with paper. However, as described above, time pressure has been suggested to be a factor that increases cognitive load, as it consumes mental resources required for performing effectively on cognitive tasks (e.g., Barrouillet et al., 2007; Burgess, 2010). Higher mental load might interact with media and generate screen inferiority, regardless of processing depth. Thus, in Experiment 2 we examined whether screen inferiority would generalize to another task-inherent cue that legitimates shallow processing but does not impose extraneous mental load—namely, low perceived importance of the task.

The same problems used in Experiment 1 were used for this experiment. These problems were introduced by Ackerman et al. (2013) in the context of what we here call the *Metacognitive Transfer Paradigm* (MTP). In this paradigm, participants attempt to solve a highly challenging initial problem, read an explanation of how to solve it, and then face a transfer problem that is similar to the first one (see details in the Materials section and an example in Appendix). As part of the MTP, participants rate their confidence in their solutions to the initial and the transfer problems immediately after providing each solution.

The critical MTP characteristic for the present study was that solution explanations for the initial problems were provided immediately after the attempt to solve them, and participants knew this in advance. To examine the spontaneous mode of work in each media, we did not ascribe levels of importance to the different phases of the task. Yet, we

hypothesized that this manipulation would lead participants to deduce that these problems were a preliminary phase, and that the main task was applying the explanation to the transfer problems. We expected this framing to lead participants who worked on screen to perceive shallow processing of the initial problems as legitimate, but less so on paper. Notably, in Experiment 1 the time frame was manipulated between participants, while in this experiment the importance manipulation took place within participants.

All participants solved the problems under a loose time frame, in the same manner as the parallel condition in Experiment 1, which did not generate screen inferiority. We hypothesized that the effects of the importance manipulation would be more pronounced on screen than on paper, and that as in Experiment 1, they would take the form of greater overconfidence, lower efficiency, and lower success rates on screen compared with paper.

3.1. Method

3.1.1. Participants

Seventy-two undergraduate students from the same population as in Experiment 1 were randomly assigned to work on screen or on paper (N = 34 and 38 per group; 53% females).

3.1.2. Materials

The six sets taken from Ackerman et al. (2013, Experiment 2) and used in Experiment 1 were used in the present experiment. Each included an initial problem, an explanation of how to solve it, and a transfer problem (see example in the Appendix). There were 77, 100, and 96 words on average in the initial problems, explanations, and transfer problems, respectively.

3.1.3. Procedure

The setting and procedure were similar to those of Experiment 1, but adapted for the three phases as in Ackerman et al. (2013). All participants were instructed as to the entire procedure in advance.

For the screen group, clicking a "Start solving" button brought up the initial problem, with an empty space below for entering the solution. After entering the solution the confidence rating scale (0-100%) was displayed. The following screen presented the problem title and the solution explanation. When done reading, a comprehension rating scale appeared, which looked like the confidence scale. On the third screen the transfer problem appeared, very much like the initial problem. For both the initial and transfer problems, participants were also asked whether they had known the problem in advance (yes/no). Also, as in Experiment 1, participants had blank sheets of paper and pens at their station for scribbling or sketching.

For the paper group, each phase of the procedure was presented on a separate page. The pages for the initial and transfer problems included space for scribbling and for writing the answer. The pages were prearranged in a pile for each participant, upside down. As in Experiment 1, participants picked up one page at a time, turning each finished page face down in a second pile. Time was documented for each phase much as in Experiment 1—i.e., via the "Start solving," "Continue," and "Next" buttons on the screen, which was otherwise empty for the paper group.

The participants had an hour to solve the entire problem set. The instructions included an explicit statement that this time allowed relaxed work, but that participants should keep

track of the time and ensure they completed the full problem set. Time reminders were announced at 30 and 50 minutes, with a final warning at 59 minutes.

3.2. Results and discussion

The effects of the medium are summarized in Table 1. Fifteen problem sets for which the initial or transfer problems were marked as known in advance (3% of the total, one per participant) were removed from the analyses. An ANOVA of the effects of the Medium (screen vs. paper) and Phase (initial vs. transfer) on the number of solutions in which participants used sketches while solving revealed two main effects and a statistically significant interaction—for the medium, $F(1,70) = 17.0$, $MSE = 957$, $p < .001$, $\eta_p^2 = .195$; for the phase, $F(1,70) = 5.40$, $MSE = 183.2$, $p = .023$, $\eta_p^2 = .072$; and for the interaction effect, $F(1,70) = 5.40$, $MSE = 184.2$, $p = .023$, $\eta_p^2 = .072$. The participants used sketches in 15% ($SD = 21.6$) of the initial problems solved on screen and in 41% ($SD = 27.9$) of those solved on paper. There was also less use of sketching when solving the transfer problems on screen ($M = 15\%$, $SD = 19.1$) compared with on paper ($M = 31\%$, $SD = 25.3$). The interaction effect stemmed from the lack of difference between the phases on screen, $t < 1$, while on paper the difference between the phases was statistically significant, $t(37) = 3.09$, $p = .004$. Thus, only the paper group showed a difference similar to that found in Experiment 1, when comparing between TP and LTF.

In this experiment, there was vast opportunity for regulation of time. Comparing the medium groups in the time they invested in the three phases revealed a main effect of the medium, with less time invested on screen than on paper, $F(1,70) = 12.73$, $MSE = 0.99$, $p = .001$, $\eta_p^2 = .154$. See Figure 2. There was also a difference between the phases, $F(2, 140) = 245.34$, $MSE = 0.92$, $p < .001$, $\eta_p^2 = .778$, stemming from shorter time invested in reading the

explanations than in solving the initial and transfer problems, which did not differ, $t(71) = 1.45$, $p = .15$. The interaction effect was not statistically significant, $F(2, 140) = 1.60$, $MSE = 0.92$, $p = .21$, $\eta_p^2 = .022$, suggesting that participants invested less time on screen than on paper during all three phases. Notably, the shorter time invested on screen is associated with the reduced use of sketches compared with the paper group. The following analyses examine whether these medium effects are associated with metacognitive monitoring and/or success rates.

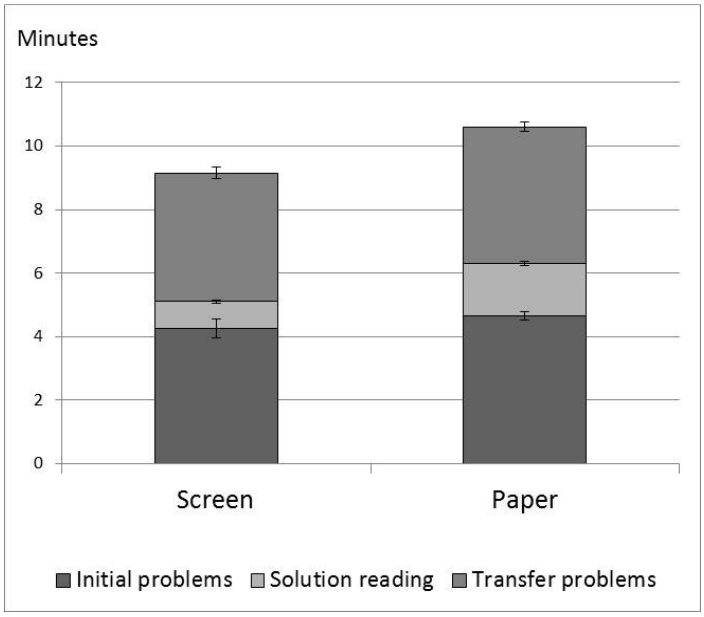

Figure 2. Experiment 2: Aggregated time investment in the three phases for screen and for paper. Error bars represent standard errors of the mean for the bar below them.

As intended, the overall success rate in the initial problems was low ($M = 18.05\%$, $SD = 17.6$), but success rates improved in the transfer problems ($M = 31.62\%$, $SD = 20.5$). See Figure 3. An ANOVA of Medium (screen vs. paper) × Phase (initial vs. transfer) revealed no main effect of the medium, $F < 1$, a strong main effect of phase, $F(1, 70) = 27.88$, $MSE = 247.13$, $p < .001$, $\eta_p^2 = .28$, and an interaction effect, $F(1, 70) = 4.02$, $MSE =$

247.13, $p = .049$, $\eta_p^2 = .054$. The interaction stemmed from greater improvement from the initial ($M = 13.7$, $SD = 13.9$) to the transfer problems ($M = 32.8$, $SD = 17.1$) on screen, $t(33) = 6.06$, $p < .001$, $d = 1.04$, compared with paper, $t(37) = 2.11$, $p = .042$, $d = 0.34$ ($Minitial = 21.9$, $SD = 19.8$; $Mtransfer = 30.5$, $SD = 23.2$). Comparing the media within each phase revealed screen inferiority only in the initial problems. Success rates in the initial problems were lower on screen ($M = 13.7$, $SD = 13.9$) than on paper ($M = 21.9$, $SD = 19.8$), $t(70) = 2.01$, $p = .044$, $d = 0.48$, while there was no difference between the two in the transfer problems, $t < 1$. Thus, the pattern of success rate differences found in Experiment 1 with and without time pressure was replicated here without time pressure, by presenting the same problems as an initial—and therefore presumably less important—phase before the main task.

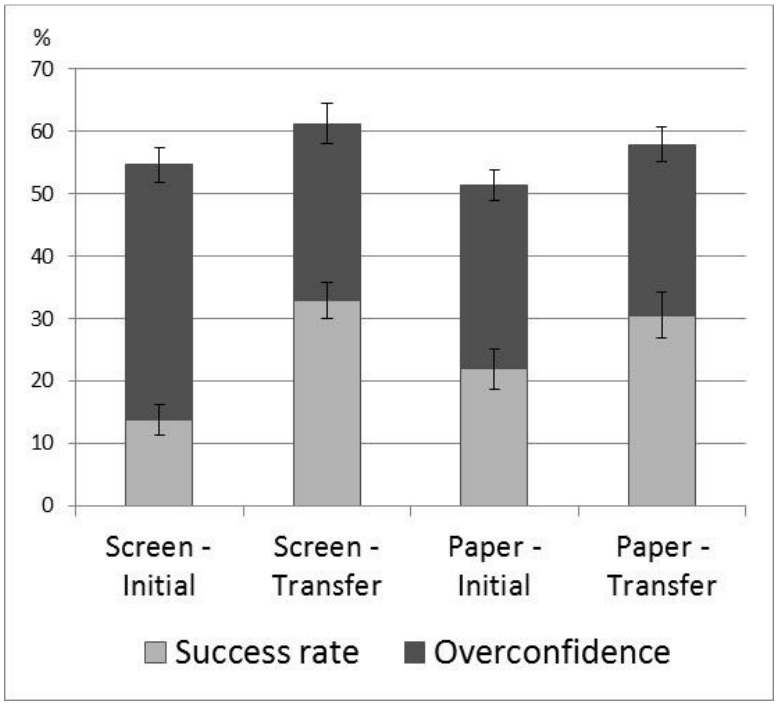

Figure 3. Success rates and overconfidence in initial and transfer solutions in Experiment 2. Confidence is represented by the top of the overconfidence bars. Error bars represent standard errors of the mean for the bar below them.

As evident in Figure 3, a pronounced overconfidence was found in all conditions (all *p*s < .001). An ANOVA on confidence revealed only a main effect of the phase, $F(1, 70) = 6.33$, $MSE = 248.37$, $p = .014$, $\eta_p^2 = .083$, reflecting lower confidence in the initial solutions ($M = 52.8$, $SD = 17.25$) than in the transfer solutions ($M = 59.5$, $SD = 20.87$), but no effect for the medium or an interaction effect, both *F*s < 1.

As in Experiment 1, confidence did not necessarily correspond with success rates. A parallel analysis on overconfidence clarifies the differential effects of phase and medium on the correspondence between confidence and success rates. This ANOVA revealed a main effect of the medium, $F(1, 70) = 4.38$, $MSE = 327.65$, $p = .04$, $\eta_p^2 = .059$, with greater overconfidence on screen ($M = 34.7$, $SD = 2.2$) than on paper ($M = 28.4$, $SD = 2.1$). There was also a main effect of the phase, $F(1, 70) = 8.18$, $MSE = 230.12$, $p = .006$, $\eta_p^2 = .11$, pointing to participants' attenuated overconfidence in their solutions to the transfer problems—a result which highlights the failure of participants' confidence ratings to rise in keeping with the actual extent of their improvement after reading the explanations. However, there was also an interaction effect, $F(1, 70) = 3.96$, $MSE = 261.31$, $p = .042$, $\eta_p^2 = .058$. This stemmed from the greater overconfidence seen on screen ($M = 40.9$, $SD = 15.9$) than on paper ($M = 29.4$, $SD = 15$) in the initial problems, $t(70) = 3.17$, $p = .002$, $d = 0.76$ with no difference between the media for the transfer problems, $t < 1$. Thus, overconfidence was greater on screen in the initial problems compared with all other conditions, as was the case under time pressure in Experiment 1.

The overall shorter time spent working on screen compared with on paper (Figure 2) may hint at shallower processing on screen throughout the process. However, success rates in

the transfer problems were parallel in both media. To calculate efficiency, we added the time invested in studying the explanations to the time spent on the transfer problems, because it is impossible to distinguish between the contributions of these two phases to performance in the transfer problems. An ANOVA as above on efficiency revealed no main effect of the medium, $F < 1$, a main effect of phase, $F(1, 70) = 16.07$, $MSE = 3.60$, $p < .001$, $\eta_p^2 = .19$, stemming from better efficiency in the transfer phase than in the initial phase, and an interaction effect, $F(1, 70) = 9.92$, $MSE = 3.60$, $p = .002$, $\eta_p^2 = .124$. In the initial problems, efficiency was not statistically different in both media, $t < 1$, suggesting that the lower success rate on screen resulted from a premature decision to stop investing effort, most likely due to overconfidence, rather than less efficient work. Notably, in the transfer phase, efficiency was marginally better on screen ($M = 4.2$, $SD = 2.2$) than on paper ($M = 3.1$, $SD = 2.4$), $t(70) = 1.97$, $p = .053$, $d = 0.47$. This high efficiency on screen was achieved despite minimal use of sketches and with less time invested than on paper.

In sum, using the MTP with no time constraints exposed that even when reducing extraneous cognitive load, the screen group was less successful and more overconfident than the paper group in the initial problems, despite the fact that they showed marginally better efficiency and similar success rates in the transfer problems. Thus, the screen group benefited from studying the explanations more than the paper group. The finding of no medium effects on solving the transfer problems suggests that screen inferiority is not inevitable, and further supports the explanation that screen performance is more susceptible to task characteristics.

4. Experiment 3

In Experiment 1 and Experiment 2, we found that time pressure and framing problems as a preliminary phase of the task generated screen inferiority in terms of metacognitive monitoring and success rates. However, these tasks still involved some reading comprehension, which is a complex multi-level process (Kintsch, 1998) that may be affected by characteristics of the presentation medium. In Experiment 3 we examined whether these results generalize even when using a challenging task that involves reading only a few isolated words, without higher-order text comprehension. We used the compound remote associates (CRA) task, which involves reading only three separated words. The task is to find a fourth word which forms a compound word or two-word phrase with each word separately. For example, for the triplet PINE/CRAB/SAUCE the correct solution is APPLE (for additional examples, see Bowden & Jung-Beeman, 2003). These problems are considered insight problems, although a recent analysis suggests that insightful solving of these problems involves the same mechanisms as involved in non-insightful solving, including working memory and attention (Chein & Weisberg, 2014). Hypothesizing that time pressure cues shallower processing on screen compared with both time pressure on paper and a loose time frame on screen regardless of the reading burden involved, we predicted a replication of Experiment 1's results and those of previous reading comprehension studies.

4.1. Method

4.1.1. Participants

One hundred and thirteen undergraduates (51% females) were randomly assigned to work on screen or on paper and to a pressured or loose time condition (26-30 participants per group).

4.1.2. Materials

The 34 CRA problems used by Ackerman and Zalmanov (2012) were used here as well. Two of the problems were used for demonstration, and the first two problems within the time frame were for self-practice.

4.1.3. Procedure

The printed instruction booklet informed participants that they would face 34 problems of varying difficulty, and detailed the procedure for each problem. On both media, the solving started when participants pressed a "Start" button on an empty screen. For the screen group this brought up a problem, with the three words presented on one line and a designated space for the solution below them. Participants provided their confidence rating in the same manner as in the previous experiments. For the paper group, each problem was printed on a separate page with its confidence rating scale. The general procedure was identical to that of the previous experiments, and the problems were randomly ordered for each participant.

All participants were invited for a 30-minute session. The actual time frames for the 32 problems that followed the first two practice problems were set by pretesting. In the pretest, participants (N = 30) were instructed to solve the problems as fast as they could, with no external time frame. This procedure resulted in a mean of 35 seconds (SD = 7.4) per problem (about 19 minutes in total). In light of this finding, the time frame in the present study was set at 16 minutes for the time pressure condition and 24 minutes for the loose time condition. The time pressure group was explicitly informed that the task was to solve the problems under time pressure and that they should allow about half a minute for each problem. It was also emphasized that they were expected to complete all the problems, despite the short time frame. The experimenter informed the participants when 5, 10, and 15 minutes had elapsed.

The loose time frame group was informed that the allotted time should allow them to work at ease, but that they should still keep track of the time and ensure they finished the entire task. The elapsed time was announced at 10, 20, and 23 minutes.

4.2. Results and Discussion

The medium effects are summarized in Table 1, together with the results of the previous experiments. The participants provided meaningful solution words (rather than answers like 'xxx' or 'don't know') for 99% of the problems, indicating that they sincerely attempted to solve the problems.

Over all analyses there were no effects of the medium, with a single exception—overconfidence. A two-way ANOVA of Medium (screen vs. paper) x Time Frame (TP vs. LTF) on solving time revealed no medium effects, Fs < 1, except for the trivial shorter time under TP (M = 27.5 seconds, SD = 3.73) compared to LTF (M = 40.4 seconds, SD = 5.76), $F(1, 109) = 200.36$, $MSE = 23.41$, $p < .001$, $\eta_p^2 = .65$. This finding replicates the pattern found in Experiment 1 with a similar manipulation but a different task.

To examine the main research question, we performed three two-way ANOVAs on success rates, confidence, and overconfidence. The results are displayed in Figure 4. For success rates, only a main effect of the time frame was found, $F(1, 109) = 7.67$, $MSE = 145.68$, $p = .007$, $\eta_p^2 = .07$, indicating that TP harmed performance (M = 45.8, SD = 13.7) compared to LTF (M = 52, SD = 9.7). There was no main effect for medium or an interaction effect, F's < 1. Confidence showed a highly similar pattern of results, $F(1, 109) = 8.54$, $MSE = 148.36$, $p = .004$, $\eta_p^2 = .07$ for time frame, with lower confidence under TP (M = 64, SD = 10.1) than under LTF (M = 57.2, SD = 13.7).

There were no main effects on overconfidence, *F*s < 1. However, there was an interaction effect, $F(1, 109) = 5.52$, $MSE = 109.72$, $p = .021$, $\eta_p^2 = .05$. A comparison between the media with respect to overconfidence showed greater overconfidence on screen ($M = 13.9$, $SD = 11.8$) than on paper under TP ($M = 8.6$, $SD = 9.6$), $t(57) = 2.21$, $p = .032$, $d = 0.50$, while under LTF, overconfidence did not significantly differ between the two media, $t(52) = 1.14$, $p = .26$, $d = 0.33$. Importantly, these effects stemmed from the fact that screen participants showed no difference for confidence between the time frames, $t(55) = 1.22$, $p = .23$, $d = 0.36$, despite having less success under TP ($M = 44.2$, $SD = 13$) than under LTF ($M = 52$, $SD = 9.6$), $t(55) = 2.65$, $p = .01$, $d = 0.69$. For paper participants, in contrast, this pattern was reversed: lower confidence under TP ($M = 56.2$, $SD = 14.6$) compared to LTF ($M = 65.7$, $SD = 8.8$), $t(54) = 2.89$, $p = .006$, $d = 0.79$, with no statistically significant success rate difference, $t(54) = 1.33$, $p = .19$, $d = 0.36$.

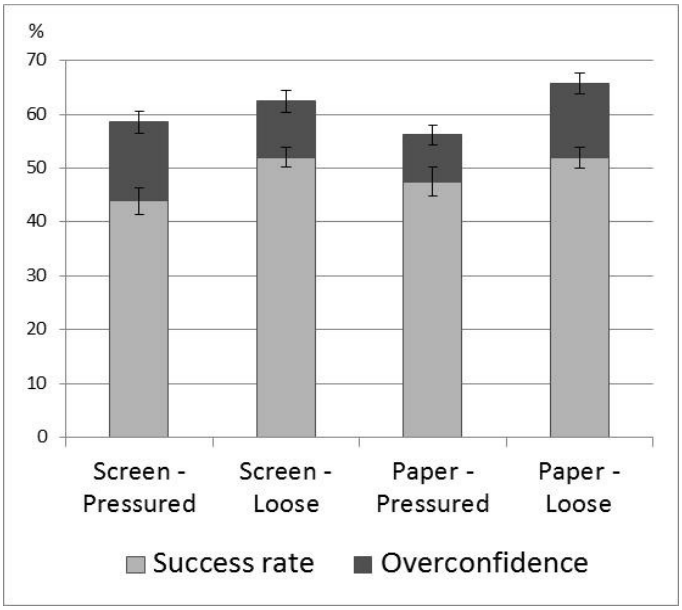

Figure 4. Success rates and overconfidence in solutions in Experiment 3. Confidence is represented by the top of the overconfidence bars. Error bars represent standard errors of the mean.

Finally, an ANOVA on efficiency revealed a main effect only for the time frame, as participants worked more efficiently under TP (M = 1.7, SD = 0.6) compared to LTF (M = 1.3, SD = 0.3), $F(1, 109) = 19.81$, $MSE = 0.22$, $p < .001$, $\eta_p^2 = .15$. Unlike in the previous experiments, there was no main effect of the medium and no interaction effect, both Fs < 1.

In sum, as expected, we found greater overconfidence on screen than on paper under time pressure even with stimuli that entailed reading only three isolated words, although no efficiency or success rate differences were found. It is evident that regardless of the reading burden, the reliability of the monitoring process is consistently more affected by task characteristics on screen than on paper.

5. General Discussion

In the present study we aimed to identify causes for screen inferiority in challenging tasks that require self-regulated effort investment, while minimizing confounding effects of reading burden, high-order reading comprehension, and cognitive load. To accomplish this, we conducted three experiments in which participants faced briefly phrased problems in either a computerized environment or a paper environment. This allowed us to expose conditions that generate screen inferiority, as detailed below. Overall, the study illuminates the medium, time pressure, and importance framing as factors affecting metacognitive monitoring, effort regulation, work efficiency, and performance.

5.1. Disentangling factors that account for medium effects on metacognitive and cognitive processes

As described above (see also Sidi et al., 2016, for a review), researchers have previously maintained that extensive reading on screen is associated with technology-related barriers,

and offered this as an explanation for screen inferiority. However, accumulated evidence raised the possibility that regulatory processes may serve as an alternative explanation. In all experiments in the present study, conditions which were expected to encourage in-depth processing, namely a loose time frame and higher perceived importance, yielded no screen inferiority—or even screen superiority—in monitoring accuracy, efficiency, and/or success rates (see Table 1). Thus, effective task performance on screen is certainly possible. Nevertheless, screen inferiority remained in the presence of task characteristics that we hypothesized to legitimate shallow processing, even with absolutely minimal reading burden.

In Experiment 1, we replicated with briefly phrased problems Ackerman and Lauterman's (2012) findings with lengthy texts. The finding of lower success rates under time pressure than under a loose time frame is consistent with a stream of the reasoning literature which associates time pressure with less-effective cognitive processing (Evans & Curtis-Holmes, 2005; Evans et al., 2009). However, the time pressure effect was only evident for the screen group, while the paper group did not compromise. Indeed, the paper group even demonstrated improved efficiency in the face of time pressure. These findings have two important implications. First, they should set off alarm bells for the research community, in that the medium in which studies were conducted (screen or paper) may turn out to have had hitherto unremarked effects on at least some known findings. Second, they testify to the possibility of effective self-regulation when conditions allow it (as was the case with the paper group in our experiment). These findings can be added to those pointing to conditions that allow highly effective regulation of reading comprehension and strategic problem solving under time pressure (Ackerman & Lauterman, 2012; Gerjets & Scheiter, 2003; Lauterman & Ackerman, 2014). A notable finding in Experiment 1 was the superiority

of working on screen under the ample time condition in terms of success and efficiency. However, as this finding was limited to one condition in Experiment 1, future research is required to draw conclusions regarding the specific conditions under which working on screen can actually benefit problem solvers.

In Experiment 2, we used the MTP procedure adapted from Ackerman et al. (2013) to generalize the conditions that lead to screen inferiority while reducing the cognitive load associated with time pressure (e.g. Barrouillet, Bernardin, Portrat, Vergauwe, & Camos, 2007; Paas & Van Merriënboer, 1994). The same task that did not generate screen inferiority under a loose time frame in Experiment 1 was now framed as a preliminary step in a sequence. We hypothesized that this change in framing would legitimate shallow processing. While we did not explicitly measure the perceived importance of the two task phases nor cognitive load, the findings demonstrate the distinct effect of reframing the task on screen versus paper. Namely, results in a within-participant design were screen inferiority in overconfidence and success rates in the initial problems, but not in the transfer problems.

There are some procedural considerations to note regarding this experiment. First, the MTP procedure does not allow counterbalancing the problems. Specifically, the "initial" problems always appear prior to the "transfer" problems due the task's nature. One could argue that the transfer problems may not have yielded screen inferiority if presented first. However, the main finding regards the comparison between the framings of the same problems as the main task in Experiment 1 and as a preliminary phase in Experiment 2. Second, in Experiment 2 participants invested less time on screen than on paper in all phases. Future research is called to further investigate the conditions that generate medium effects on regulation of time, in addition to effects on monitoring, efficiency, and outcomes. Finally,

the importance manipulation may have affected participants' motivation. Specifically, perhaps perceiving the transfer problems as the more important aspect of the task, raised the motivation to succeed in them. Thus, it is possible that increased motivation for success may be a moderating factor for screen inferiority. Future research is required to determine the contribution of motivation to processing depth on screen.

In Experiment 3, we eliminated high-order reading comprehension altogether by using the CRA problems. Unlike in Experiment 1 and Experiment 2 which involved reading comprehension, here performance and efficiency were not affected by the medium, suggesting on a possible interaction with text length or higher-order processing. Importantly though, there was still greater overconfidence on screen than on paper under time pressure, but not under a loose time frame. In addition, there was lower sensitivity of confidence ratings to performance variations between the time conditions, as found before with font readability (Sidi et al., 2016). It may be argued that reaching similar efficiency and performance for both media is satisfactory, even if a monitoring bias remains. However, monitoring biases are profoundly problematic, since they are expected to misguide future regulatory decisions (e.g., Metcalfe & Finn, 2008).

Overall, the results support our hypothesis that working on screen is highly sensitive to task characteristics that signal legitimacy for shallow processing, and this affects both metacognitive and cognitive processes. On paper, in contrast, the default mode of work is characterized by in-depth processing, even in the presence of such task characteristics. One possible explanation for this stems from the different type of interactions that characterize these media. Namely, the typical interactions on screen involve brief reading of e-mails, social networking posts, forums, etc. This daily computerized interaction promotes

differentiated reading behavior from paper, with more selective reading, browsing, and scanning (Liu, 2005; Mizrachi, 2015). Another possible explanation relates to how people perceive the two media. Going long back, considering television screens versus paper-based sources of information, Salomon (1984) suggested that the mental effort invested while learning from any medium depends on its perceptions of ease and seriousness. Despite the difference in the electronical modality we consider, this explanation seems relevant to computerized environments as well and accords our interpretation of the conditions that lead to screen inferiority.

Our proposed account of conditions under which it is possible to eliminate screen inferiority can shed light on previous findings. For instance, Eden and Eshet-Alkalai (2013) described as surprising their finding of equivalence between the media in a text editing task. In light of the present study, it appears that text editing may trigger in-depth processing, explaining the found equivalence. To take another example, Norman and Furnes (2016) replicated the limited time frame condition used in Ackerman and Lauterman (2012) and found equivalence between the media in both overconfidence and performance, rather than screen inferiority. However, they did not make it clear whether the time frame used was pressured or loose for the given task in their population.[2] According to the present study, only if the time frame was perceived by the participants as pressured would we expect it to hint at legitimacy for shallower processing on screen and generate screen inferiority.

[2] Other methodological differences may also account for the discrepancy between the findings of Norman and Furnes (2016) and Ackerman and Lauterman (2012). In the former, (1) multiple and repeated judgments were measured; (2) both study media were used in the same session, in a within-participant design which may generate awareness to the media and affect the results; (3) memory for details was measured but not higher-order comprehension; (4) the analyses were done after controlling for subjectively reported effort despite its strong relevant to self-regulation; and (5) any medium preference in the population was unknown.

5.2. Theoretical implications

Combined, our findings allow theoretical analysis of factors that affect metacognitive processes in general, and in meta-reasoning, in particular. This broader contribution is especially important for the nascent field of meta-reasoning research, since not much is known about the factors which affect monitoring and effort regulation in this context (see Ackerman & Thompson, 2015).

In presenting the meta-reasoning framework, Ackerman and Thompson (2015) attempted to draw both parallels and distinctions between meta-memory (metacognitive aspects of memorizing word lists), meta-comprehension (metacognitive aspects of reading comprehension tasks), and meta-reasoning processes (metacognitive aspects of problem solving). Several empirical studies have already identified processes which differ among these domains (Ackerman, 2014; Thompson et al., 2013). The present study, like previous work in this research line, shows highly similar patterns of results for meta-comprehension and meta-reasoning processes.

A close look at Table 1 reveals that in all three experiments, confidence ratings were blind to medium effects on performance. If the medium were the only manipulated factor, we could conclude that confidence is an insensitive measure which fails to reflect variations in performance. However, in most cases, confidence ratings were sensitive to the other factors we manipulated. Confidence varied with perceived importance of the task, in a within-participant design (Experiment 2), but also with time frame, in a between-participants design (Experiment 1 in both media and Experiment 3 on paper, but not on screen). Previous studies point to higher sensitivity of monitoring to within-participant variations than to variations between groups (e.g., Koriat, Ackerman, Adiv, Lockl, & Schneider, 2014). We found

confidence to be particularly sensitive to our manipulations, but not to differences between the media. These findings are in line with the cue utilization approach to metacognitive monitoring (Koriat, 1997), which suggests that correspondence between monitoring and performance stems from utilization of heuristic cues for confidence which reliably reflect variations in performance (see Koriat, 2008). The metacognitive literature often highlights that monitoring is more sensitive to experience-based cues derived from processing each item than to cues external to the itemized task, like repeated memorizing of the same list (Koriat, Sheffer, & Ma'ayan, 2002). The present line of research adds the work environment to the set of external cues that are not adequately taken into account when monitoring one's likelihood of success.

The association between monitoring reliability and depth of processing found in the present study suggests that under proper conditions certain cues can trigger deeper processing than people tend to engage in spontaneously. To our knowledge, all previous empirical evidence for this association is in the domain of reading comprehension (see Thiede, Griffin, Wiley, & Anderson, 2010, for a review). Notably, most of this previous research dealt with a different aspect of monitoring reliability than the present study, namely, resolution. Resolution, or relative accuracy, is a measure of the extent to which monitoring of knowledge discriminates between better and less well-known items.

The present study deals with calibration, or absolute accuracy, where mean confidence ratings are compared with success rates on the complete task; in the example above, calibration would be good if the student expected to get about 80 percent of the questions correct across the entire exam, and did so. Overconfidence thus reflects poor calibration. Lauterman and Ackerman (2014) presented evidence that calibration could be

improved (i.e., overconfidence reduced) on screen using a study strategy that had previously been found to improve resolution—namely, writing keywords summarizing the gist of a given text after studying it. The present study adds that cues inherent in the task can also affect calibration, but that this improvement depends on the work medium. These findings are intriguing in several respects. First, as mentioned above, the present findings represent the first association between depth of processing and monitoring reliability observed in the context of problem solving. Second, the finding that cues inherent in the task may improve calibration raises the question of whether this is also true in reading comprehension tasks, possibly providing another source of commonality between meta-comprehension and meta-reasoning processes. Finally, the present findings bolster the idea that the medium is a consistent cue which interacts with other cues to generate effects on metacognitive processes. All these issues deserve further research aimed at extending our understanding of the involved metacognitive processes.

In conclusion, the present study emphasizes the importance of distinguishing cues that legitimate shallow processing from those which trigger in-depth processing. More broadly, we highlight the susceptibility of metacognitive processes to contextual cues.

5.3. Practical implications

The shift from paper-based to computer-based work is obviously unavoidable. In light of this fact, the apparent persistence of screen inferiority despite the most recent technological advances is troubling. As such, it behooves us to give deep thought to possible effects of the medium on common daily-life and educational tasks, many of which have real consequences for the individuals and institutions involved (e.g., work and educational screening exams). Indeed, our study highlights conditions that allow avoiding screen

inferiority. We demonstrated that using task-inherent cues which call for depth of processing, or avoiding those that legitimate shallow processing, may make the difference between perpetuating screen inferiority and overcoming it, or even achieving screen superiority.

The present findings suggest implications for designing computerized environments for learning, assessment, and daily tasks. Most important, designers should take into account the types of task characteristics that might result in inferior performance. One such example is time pressure. Many testing environments operate on the basis of strict time frames, including SAT tests (deDonno, Rivera-Torres, Monis, & Fagan, 2014), GED tests (GED Testing Service, 2002) and some MOOCS (e.g., Mœglin & Vidal, 2015; Severance, 2013). Following our findings, time pressure in these digital settings might produce biased assessments of participants' abilities. The good news is that supplementing the task with cues that support in-depth processing can encourage participants to engage in effective processing even in computerized environments. The potential association we considered between cognitive load and time pressure may point to developing additional methods for eliminating screen inferiority, based on methods developed for reducing cognitive load, in particular those found effective in the context of e-learning (see Kirschner, Ayres, & Chandler, 2011 and van Merriënboer & Ayres, 2005, for reviews).

At the global level, the present study is part of an endeavor to increase the flow of knowledge from experimental cognitive psychology into educational research (de Bruin & van Gog, 2012). It has already been shown that applying insights gleaned from experimental metacognitive studies to educational contexts, although sometimes challenging, is feasible (e.g., Baars, Vink, van Gog, de Bruin, & Paas, 2014; Metcalfe, Kornell, & Son, 2007; Redford, Thiede, Wiley, & Griffin, 2012; Roebers, Schmid, & Roderer, 2009; van Loon, de

Bruin, van Gog, van Merriënboer, & Dunlosky, 2014). We designed the present study to be as close as possible to educational settings, by using task types and environments that are common in such contexts, without losing the advantages of the well-controlled laboratory setting. This should contribute to the relevance of the findings with respect to appropriate conditions for effective computerized work and cues for depth of processing. We hope that educational researchers will continue our cognitive research and pave the way for applying the insights it has yielded in classrooms.

Taking another perspective, the present study may shed light on cases in which students struggle with demanding tasks, and suggest possible strategies for improving their effort regulation. For instance, it is well-established that school students find verbal math problems highly challenging (e.g., Morsanyi et al., 2014; Múñez, Orrantia & Rosales, 2013). It is possible that many situations in which students encounter such problems involve computerized environments and time pressure, and/or a framing of problems as training toward a future exam. This insight may lead educators to adjust the learning environment so as to provide cues that hint at the importance of the task and avoid those which hint at legitimacy for shallow processing. Taking this even further, it is well-established that up-to-date pedagogy needs to be adjusted to computerized environments, and that tasks cannot simply be transferred from traditional study environments to computerized ones (Angeli & Valanides, 2009; Mishra & Koehler, 2006). However, there are no clear guidelines as to how to do this effectively (see Cheung & Slavin, 2013, for a review). We hope that the conclusions from the present study regarding cues for depth of processing will inspire the development of pedagogical guidelines for effective computerized learning.

References

Ackerman, R. (2014). The diminishing criterion model for metacognitive regulation of time investment. *Journal of Experimental Psychology: General*, *143*(3), 1349–1368.

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

Ackerman, R. & Thompson, V. (2015). Meta-Reasoning: What can we learn from meta-memory. In A. Feeney & V. Thompson (Eds.), *Reasoning as Memory* (pp. 164-178). Hove, UK: Psychology Press.

Ackerman, R., & Zalmanov, H. (2012). The persistence of the fluency–confidence association in problem solving. *Psychonomic Bulletin & Review, 19*(6), 1189–1192.

Angeli, C., & Valanides, N. (2009). Epistemological and methodological issues for the conceptualization, development, and assessment of ICT-TPCK: Advances in technological pedagogical content knowledge (TPCK). *Computers and Education, 52,* 154–168.

Antón, C., Camarero, C., & Rodríguez, J. (2013). Usefulness, enjoyment, and self-image congruence: The adoption of e-book readers. *Psychology & Marketing*, *30*(4), 372–384.

Ayres, P., & Paas, F. (2007). Making instructional animations more effective: A cognitive load approach. *Applied Cognitive Psychology*, *21*(6), 695-700.

Baars, M., Vink, S., van Gog, T., de Bruin, A., & Paas, F. (2014). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learning and Instruction, 33*, 92–107.

Ball, R., & Hourcade, J. P. (2011). Rethinking reading for age from paper and computers. *International Journal of Human-Computer Interaction, 27*(11), 1066–1082.

Baron, N. S. (2013). "But still it moves": Screens, print, and reading. *Selected Papers of Internet Research*, *3*, 1–5.

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 570-585.

Benedetto, S., Drai-Zerbib, V., Pedrotti, M., Tissier, G., & Baccino, T. (2013). E-readers and visual fatigue. *PloS One*, *8*(12), 1–7.

Ben-Yehudah, G., & Eshet-Alkalai, Y. (2014). The influence of text annotation tools on print and digital reading comprehension. In Y. Eshet, A. Caspi, N. Geri, Y. Kalman, V. Silber-Varod, & Y. Yair (Eds.), *Proceedings of the 9th Chais Conference for Innovation in Learning Technologies* (pp. 28–35). Raanana, Israel: Open University Press.

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

Bowden, E. M., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods, Instruments, & Computers*, *35*(4), 634–639.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Burgess, D. J. (2010). Are providers more likely to contribute to healthcare disparities under high levels of cognitive load? How features of the healthcare setting may lead to biases in medical decision making. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, *30*(2), 246-257.

Chein, J. M., & Weisberg, R. W. (2014). Working memory and insight in verbal problems: Analysis of compound remote associates. *Memory & Cognition*, *42*(1), 67–83.

Cheung, A., & Slavin, R. E. (2013). The effectiveness of educational technology applications on mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review, 9*, 88–113.

Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 143–230). Dordrecht, Netherlands: Springer.

Daniel, D. B., & Woody, W. D. (2013). E-textbooks at what cost? Performance and use of electronic vs. print texts. *Computers & Education, 62*, 18–23.

de Bruin, A. B., & van Gog, T. (2012). Improving self-monitoring and self-regulation: From cognitive psychology to the classroom. *Learning and Instruction, 22*(4), 245–252.

deDonno, M. A., Rivera-Torres, K., Monis, A., & Fagan, J. F. (2014). The influence of a time limit and bilingualism on Scholastic Assessment Test performance. *North American Journal of Psychology*, *16*(2), 211–224.

DeStefano, D., & LeFevre, J. A. (2007). Cognitive load in hypertext reading: A review. *Computers in human behavior*, *23*(3), 1616-1641.

Dennis, A. R., Abaci, S., Morrone, A. S., Plaskoff, J., & McNamara, K. O. (2016). Effects of e-textbook instructor annotations on learner performance. *Journal of Computing in Higher Education, 28*(2), 221-235..

Eden, S., & Eshet-Alkalai, Y. (2013). The effect of format on performance: Editing text in print versus digital formats. *British Journal of Educational Technology*, *44*(5), 846–856.

Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review, 13*(3), 378–395.

Evans, J. St. B. T. & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, *11*(4), 382–389.

Evans, J. St. B. T., Handley, S. J., & Bacon, A. M. (2009). Reasoning under time pressure: A study of causal conditional inference. *Experimental Psychology*, *56*(2), 77–83.

GED Testing Service. (2002). Who took the GED? GED Statistical Report. Washington, DC: American Council on Education.

Gerjets, P., & Scheiter, K. (2003). Goal configurations and processing strategies as moderators between instructional design and cognitive load: Evidence from hypertext-based instruction. *Educational Psychologist*, *38*(1), 33-41.

Gu, X., Wu, B., & Xu, X. (2015). Design, development, and learning in e-textbooks: What we learned and where we are going. *Journal of Computers in Education, 2*(1), 25–41.

Hillesund, T. (2010). Digital reading spaces: How expert readers handle books, the Web and electronic paper. *First Monday (Online), 15*(4). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2762/2504.

Höffler, T. N., & Leutner, D. (2007). Instructional animation versus static pictures: A meta-analysis. *Learning and instruction*, *17*(6), 722-738.

Hollender, N., Hofmann, C., Deneke, M., & Schmitz, B. (2010). Integrating cognitive load theory and concepts of human–computer interaction. *Computers in Human Behavior*, *26*(6), 1278-1288.

Holzinger, A., Baernthaler, M., Pammer, W., Katz, H., Bjelic-Radisic, V., & Ziefle, M. (2011). Investigating paper vs. screen in real-life hospital workflows: Performance contradicts perceived superiority of paper in the user experience. *International Journal of Human-Computer Studies*, *69*(9), 563–570.

Kazanci, Z. (2015). University students' preferences of reading from a printed paper or a digital screen—A longitudinal study. *International Journal of Culture and History, 1*(1), 50–53.

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

Kirschner, P. A., Ayres, P., & Chandler, P. (2011). Contemporary cognitive load theory research: The good, the bad and the ugly. *Computers in Human Behavior*, *27*(1), 99–105.

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

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(4), 945–959.

Koriat, A., Ackerman, R., Adiv, S., Lockl, K., & Schneider, W. (2014). The effects of goal-driven and data-driven regulation on metacognitive monitoring during learning: A developmental perspective. *Journal of Experimental Psychology: General, 143*(1), 386–403.

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.

Kretzschmar, F., Pleimling, D., Hosemann, J., Füssel, S., Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2013). Subjective impressions do not mirror online reading effort: Concurrent EEG-eyetracking evidence from the reading of books and digital media. *PloS One*, *8*(2), e56178.

Kühl, T., & Eitel, A. (2016). Effects of disfluency on cognitive and metacognitive processes and outcomes. *Metacognition and Learning, 11*(1), 1-13.

Lauterman, T., & Ackerman, R. (2014). Overcoming screen inferiority in learning and calibration. *Computers in Human Behavior*, *35*, 455–463.

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing*, *6*(1), 1–24.

Lin, C. L., Wang, M. J. J., & Kang, Y. Y. (2015). The evaluation of visuospatial performance between screen and paper. *Displays*, *39*, 26–32.

Liu, Z. (2005). Reading behavior in the digital environment. *Journal of Documentation, 61*(6), 700–712.

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.

Margolin, S. J., Driscoll, C., Toland, M. J., & Kegler, J. L. (2013). E-readers, computer screens, or paper: Does reading comprehension change across media platforms? *Applied Cognitive Psychology, 27*, 512–519.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23.

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

Metcalfe, J., Kornell, N., & Son, L. K. (2007). A cognitive-science based programme to enhance study efficacy in a high and low risk setting. *European Journal of Cognitive Psychology*, *19*(4–5), 743–768.

Meyer, A., Frederick, S., Burnham, T., Guevara Pinto, J. D., Boyer, T. W., Ball, L. J., Pennycook, G., Ackerman, R., Thompson, V., & Schuldt, J. P. (2015). Disfluent fonts don't help people solve math problems. *Journal of Experimental Psychology: General, 144*(2), e16-e30.

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054.

Mizrachi, D. (2015). Undergraduates' academic reading format preferences and behaviors. *The Journal of Academic Librarianship*, *41*(3), 301–311.

Mœglin, P., & Vidal, M. (2015). Managing time, workload and costs in distance education: Findings from a literature review of Distances et Médiations des Savoirs (formerly Distances et Savoirs). *Distance Education*, *36* (2), 282–289.

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*(3), 329–348.

Morsanyi, K., Busdraghi, C., & Primi, C. (2014). Mathematical anxiety is linked to reduced cognitive reflection: a potential road from discomfort in the mathematics classroom to susceptibility to biases. *Behavioral and Brain Functions*, *10*(1), 1-13.

Moustafa, K. (2016). Improving PDF readability of scientific papers on computer screens. *Behaviour & Information Technology*, *35*(4), 319-323.

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard advantages of longhand over laptop note taking. *Psychological science, 23,* 1159–1168.

Múñez, D., Orrantia, J., & Rosales, J. (2013). The effect of external representations on compare word problems: Supporting mental model construction. *The Journal of Experimental Education, 81*(3), 337–355.

Murray, M. C., & Pérez, J. (2011). E-textbooks are coming: Are we ready? *Issues in Informing Science and Information Technology*, *8*, 49–60.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation*, *26*, 125–141.

Norman, E., & Furnes, B. (2016). The relationship between metacognitive experiences and learning: Is there a difference between digital and non-digital study media? *Computers in Human Behavior*, *54*, 301–309.

Paas, F. G., & Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational psychology review*, *6*(4), 351-371.

Prowse Turner, J. A., & Thompson, V. A. (2009). The role of training, alternative models, and logical necessity in determining confidence in syllogistic reasoning. *Thinking & Reasoning*, *15*(1), 69–100.

Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and testing. *Science*, *323*(5910), 75–79.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(3), 435–451.

Redford, J. S., Thiede, K. W., Wiley, J., & Griffin, T. D. (2012). Concept mapping improves metacomprehension accuracy among 7th graders. *Learning and Instruction, 22*(4), 262–270.

Roebers, C. M., Schmid, C., & Roderer, T. (2009). Metacognitive monitoring and control processes involved in primary school children's test performance. *British Journal of Educational Psychology, 79*(4), 749–767.

Salmerón, L., & García, V. (2012). Children's reading of printed text and hypertext with navigation overviews: The role of comprehension, sustained attention, and visuo-spatial abilities. *Journal of Educational Computing Research*, *47*(1), 33–50.

Salomon, G. (1984). Television is" easy" and print is" tough": The differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of educational psychology*, *76*(4), 647-658.

Severance, C. (2013). MOOCs: An Insider's View. *Computer, 46*(10), 93-96.

Schnotz, W., & Kürschner, C. (2007). A reconsideration of cognitive load theory. *Educational Psychology Review*, *19*(4), 469-508.

Shynkaruk, J. M., & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory & Cognition*, *34*(3), 619–632.

Sidi, Y., Ophir, Y., & Ackerman, R. (2016). Generalizing screen inferiority: Does the medium, screen versus paper, affect performance even with brief tasks? *Metacognition and Learning, 11*(1), 15-33.

Singer, L. M., & Alexander, P. A. (2017). Reading across mediums: Effects of reading digital and print texts on comprehension and calibration. *The Journal of Experimental Education*, *85*(1), 155-172.

Sweller, J. (1976). The effect of task complexity and sequence on rule learning and problem solving. *British journal of Psychology*, *67*(4), 553-558.

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational psychology review*, *10*(3), 251-296.

Thiede, K. W., Anderson, M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of educational psychology*, *95*(1), 66-73.

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 1024–1037.

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes*, *47*(4), 331–362.

Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition*, *128*(2), 237–251.

Valcke, M. (2002). Cognitive load: updating the theory?. *Learning and Instruction*, *12*(1), 147-154.

van Horne, S., Russell, J., & Schuh, K. L. (2016). The adoption of mark-up tools in an interactive e-textbook reader. *Educational Technology Research and Development, 64*(3), 407-433.

van Loon, M. H., de Bruin, A. B., van Gog, T., van Merriënboer, J. J., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica, 151*, 143–154.

van Merriënboer, J. J., & Ayres, P. (2005). Research on cognitive load theory and its design implications for e-learning. *Educational Technology Research and Development*, *53*(3), 5–13.

Winne, P. H., & Baker, R. S. (2013). The potentials of educational data mining for researching metacognition, motivation and self-regulated learning. *JEDM-Journal of Educational Data Mining*, *5*(1), 1–8.

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

Appendix - Example of the problems used in Experiment 1 and Experiment 2

The following problem was used as a main task in Experiment 1 and as an initial problem in Experiment 2:

> Joe and Dan are old friends who have not met for many years. As they catch up, Joe asks Dan how many kids he has. "Three," answers Dan. "And how old are they?" says Joe.
>
> "Well," says Dan, "the product of their ages is 36." "Hmm," says Joe, "can you give me a little more information?" "Okay," says Dan. "The sum of their ages is exactly the number of beers we had today." "That helps," says Joe, "but it's not quite enough."
>
> "Okay," says Dan. "So I'll add that the elder two have green bikes." Joe now knows how old the kids are. How?

Transfer problem used for the initial problem above in Experiment 2:

> Joe and Dan are old friends who have not met for many years. Joe asks Dan: "What are the ages of your three kids?" Dan answers, "None of the children are less than two years old, and the sum of their ages is 14." "Can you give me more information?" says Joe. So Dan adds, "The product of their ages is exactly the house number of this pub." "That's not enough," says Joe.
>
> "Okay," says Dan. "I'll add that my young twins' names are Milly and Julie." Joe now knows how old the kids are. How?