Learning and Instruction 51 (2017) 90–97

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

Cognitive load and self-regulation: Attempts to build a bridge

Monique Boekaerts

Department of Education, Leiden University, The Netherlands

A R T I C L E I N F O

Article history:
Received 18 April 2017
Received in revised form
12 June 2017
Accepted 4 July 2017
Available online 13 July 2017

Keywords:
Cognitive load
Self-regulation
Multiple measurement points
New methodologies to capture cognitive load

A B S T R A C T

The editors of the Special Issue called for a more integrative approach to the study of cognitive load and self-regulation. The goal formulated for the Special Issue is ambitious. In my role as a constructive critic, I first summarized the findings in the 6 papers, identifying important questions and concerns that emerged while reading the papers. I also identified some general issues that need further clarification and elaboration: I argued that there is a strong need to reach consensus on the conceptualization and measurement of cognitive load and that new methodologies should be developed to capture cognitive load *in real time* and link it to strategy use.

1. Introduction

The editors of the special issue argued that due to the explosion of information it is absolutely essential that students learn to filter, select, and process incoming information and that teachers learn to design instruction in such a way that their students can acquire these self-regulation strategies. In other words, 2 complementary research traditions are involved in this important skill building processes, namely self-regulated learning and instructional design. These 2 research traditions have their own histories, theories, measuring instruments, and types of interventions.

The title of the special issue reveals that the editors called for a more integrative approach to the study of self-regulation and cognitive load. They argue that research in these two separate lines of research shows little overlap, even though studies often depart from the same or similar research questions. The goal formulated for the Special Issue is ambitious. The editors invited 6 research groups and asked them (1) to present innovative, empirical research that links the two domains of research and (2) to discuss how bringing together the 2 vast bodies of research can provide the foundation for research on contemporary issues in educational science. Has this ambitious goal been accomplished? In the next sections, I will first summarize the findings of the 6 papers, identifying important questions and concerns that emerged while reading the papers. Next, I will point to areas in need of investigation.

1.1. Summaries and critical issues related to the 6 articles

The focus of 3 of the manuscripts was on depth of processing, namely the papers by Schleinschok et al., Glogger-Frey et al., and Sidi et al. The former two studies wanted to improve depth of processing through improved metacognitive regulation. Both research groups argued that students often overestimate their level of understanding of a text and that this implies that they stop short of grasping its full meaning. Each research group proposed a specific cognitive strategy that could help students to improve their self-regulation strategies and they set up experiments to demonstrate that use of this strategy would result in more efficient monitoring and control. Sidi and her co-workers addressed a related question, namely: Can depth of processing be triggered by contextual cues?

Schleinschok, Eitel, and Scheiter (2017) predicted that instructing students to make a free-hand drawing of the content of a paragraph would be instrumental to (1) more accurate monitoring that allows students to make inferences that are directly relevant to understand the deep structure of the text and (2) to better cognitive control of the quality of what they encoded in their memory schemata. They set up 2 experiments with university students. In both experiments students were split up into a drawing group and a text-only group. All students had to read an expository text. After reading each of the 5 paragraphs they had to indicate

E-mail address: Boekaerts@fsw.leidenuniv.nl.

http://dx.doi.org/10.1016/j.learninstruc.2017.07.001

how confident they were that they had encoded the paragraph well (Jol). The students in the drawing condition made a free-hand drawing after reading each paragraph. After reading the full text, they rated the quality of their encodings and indicated which paragraphs they wanted to re-study. Finally, they were requested to rate the degree of cognitive load (CL) that they had experienced. The research group postulated that drawing the content of a paragraph might be more effective than other learning strategies, such as summarizing and paraphrasing the text, because students will anticipate that they have to construct a coherent internal pictorial memory code, in addition to a verbal memory code. The researchers predicted and found that generating free-hand drawings *after* reading a paragraph leads to a more accurate metacognitive judgement of the quality of their learning (Jol) that better matched performance on the posttest. Students in the drawing condition were not only more aware of the quality of their understanding, they also preferred to re-study paragraphs for which they had the lowest Jol's, thus demonstrating that anticipation of the free-hand drawing task did not only guide and support their monitoring but it also informed them on the knowledge gaps they still had and needed to fill.

Experiment 2 was similar to experiment 1, but all students were allowed to restudy the paragraphs they had indicated for restudy. Contrary to expectations, students in both conditions selected paragraphs for re-study for which their JoL's were lower and they spent more time restudying them. In line with previous research, a high score on experienced CL was associated with lower scores on the different posttests (drawing task, verification task, and diagram labeling) in both experiments. However, this relation disappeared when JoLs were simultaneously entered into the regression model (this is not surprising given the high negative correlation (−0.69) between the two variables). The researchers concluded that JoL's rather than the experienced CL predicted post-test performance in the drawing condition. Even though I am not convinced that this research group captured CL in a valid way (see my discussion in the section on the meaning and measurement of CL), they demonstrated that requesting students to make a visual representation of the content of a paragraph *after* they finished reading it, is an active generative task that makes them aware that it is not sufficient to monitor at the surface level. Inspection of the internal pictorial code may act as a strong cue that more accurate monitoring is necessary to discover the deep structure of each paragraph.

Glogger-Frey, Gaus and Renkl (2017) set out to demonstrate that encouraging students to detect the rule or principle in multiple cases results in better understanding the deep structure of a text or problem. They designed a SRL environment and trained 8th grade students in a 20 min SRL training session to monitor for the critical features in a ratio problem. They compared the effect of the invention group with the performance of students who worked with guided examples. The two groups were compared on different process variables, such as their level of encoding, self-efficacy, awareness about knowledge gaps, and experienced CL, as well as their performance on a transfer task. The experiment was conducted in 2 regular school lessons. At the end of the first training sessions, the groups did not differ significantly on self-reports that assessed their self-efficacy and perceived knowledge gaps, but a main condition effect was noted on reported extraneous load measured with a 5 item scale. Extraneous load interacted with self-perceived performance in math and science (measured just before the training sessions), suggesting that only students who believed their math and science performance to be low had indicated that the extraneous load of the self-regulated activity was high (see my comments on capturing CL in the discussion section). The recall test conducted four days later revealed that there was no significant difference in the encoding of the surface features of the problems, but that students in the invention condition showed a deeper encoding of the problem's ratio structure. The students then worked on a second problem in their respective conditions, followed by a ten minutes lecture on ratios in physics, and by a near and far transfer task. It was predicted and found that the students who had worked in the invention condition would outperform students in the guided condition on the near and far transfer problems. Interestingly, both deep-structure encoding and extraneous load mediated the effect of the type of training sessions on transfer performance. And further exploratory analyses revealed that students in the invention condition had improved their in-depth processing during the *second* invention phase. By contrast, the explanations that the students in the guided practice conditions gave for the steps that an imaginary student took got worse during the second guided session. It is a pity that the researchers failed to measure students' self-efficacy, perception of extraneous load, awareness of knowledge gaps, and level of encoding in relation to the second practice sessions. This would have given us more insight into the advantages and disadvantages of having 2 training sessions. For example, would the students in the invention condition still report higher extraneous CL in the second invention session? Did insight into the structural relations occur already during the first SR training session in some student pairs and was it consolidated in the second training session in all student pairs? What was the exact role of the direct instruction that followed the second training sessions in consolidating this insight? Did it have the same effect in both conditions? These are but a few questions that await scientific investigation (see my discussion on multiple time points).

Sidi, Shpigelman, Zalmanov, and Ackerman (2017) informed the reader that the results of studies that compared students' performance in computerized learning environments with learning from texts in traditional environments are inconclusive. They clarified that the reported lower performance on screen as well as persistence of a paper preference in all age groups are not caused by technological disadvantages but are due to a qualitatively different reading process, characterized by many interruptions, attentional shifts, and multi-tasking. Overconfidence and less efficient work on screen contrast sharply with the reliable monitoring displayed while reading on paper. Sidi et al. (2017), hypothesized that screen environments encourage students to adopt a shallower processing style than paper environments, especially when they pick up cues that legitimize shallow processing. They wanted to know whether screen inferiority was due to the CL created by reading lengthy texts, or whether screen inferiority could also be demonstrated regardless of the reading burden. They set up 3 experiments to study the effect of different manipulations on response time, reported confidence after doing a task, calculated overconfidence, processing efficiency (correct solutions per hour), and success rate. They selected 6 challenging logic problems that could be stated briefly. In the first experiment, undergraduates were randomly allocated to the on screen and on paper group and worked either under time pressure (TP) or in a loose time frame (LTF). As predicted, TP resulted in screen inferiority reflected in lower processing efficiency and success rates, as well as poor calibration. Remarkably, when working in a LTF, these students showed more efficiency, higher success rates, and no differences in overconfidence, compared to the on paper group.

In the second experiment, the same problems were designed in a metacognitive transfer paradigm. Each of the 6 problem sets consisted of the following procedure: solving an initial problem, followed by a confidence rating, an explanation of the problem solution, solving a transfer problem, and a confidence rating. Students were told that they could work in a LTF but needed to monitor the time to complete the full problem set in time. The overall success rate of the initial problems was low, but improved in the transfer tasks. The on screen group invested less time in solving the initial and transfer problems and also devoted less time to reading the explanations of the solutions compared with the on paper group. Separate analysis of the initial problem data yielded screen inferiority (i.e., lower success rate and more overconfidence) but no significant difference in processing efficiency. When analyzing the transfer problem data separately, no significant differences were found between the on screen group and the on paper group in success rate and calibration score, but the processing efficiency of the on screen group was significantly *better* despite minimal use of sketches and less time invested. Sidi et al., interpreted these results as evidence for screen inferiority under conditions of low perceived importance. They explained that - similar to TP – low perceived importance is interpreted by the students as a sign that adopting a more shallow processing style is legitimate. I have some difficulties understanding the researchers' line of argument. They maintain that the experimental design had framed the initial problems as 'preliminary' and that the students in the on-screen condition had picked up that signal, thus appraising the initial problems as 'low in importance'. Yet, I could neither find instructions to this effect nor could I detect any evidence of a manipulation check in the manuscript. We need to know whether or not the students detected this cue and how they interpreted it, before conclusions can be drawn about the relation between this cue and the difference in processing style during the initial and the transfer problems. The third experiment is a replication of the first one, using 32 three word compound remote associate tasks. As predicted, screen inferiority in overconfidence was found under TP but not under LTF. None of the other measures differed significantly.

Sidi et al., reported important findings that have serious implications for the way we test and evaluate students. They showed that working on screen can be as effective as working on paper, but that students who work on screen are more susceptible to cues that trigger shallow processing. An avenue for future investigation would be to document the differences in cognitive and metacognitive strategies that students who work under these different conditions use. I will return to this important point shortly. Here, it will suffice to state that the meaning students attach to learning tasks, whether they are presented on screen or on paper, *unfolds* from the moment they interpret the task and the task demands till well after task completion.

Van Loon, Destan, Spiess, De Bruin and Roebers (2017) maintain that only individuals, who are able to accurately self-evaluate learning, are able to engage in adaptive self-regulation. They found that the SR strategies that young children use are often not adapted to the situation. Possible reasons are that young children cannot yet differentiate well between correct and incorrect responses and that they are overconfident that they can do the task. This implies that they often start a learning task without thinking and use the first learning strategy that comes to mind. Van Loon and colleagues wanted to know what the reasons are for young children's inaccurate self-evaluations and their overconfidence. They recruited kindergarten pupils (approximately 6 years old) and second graders (8 years old) to study age differences in overconfidence and accuracy of self-evaluations. They presented Asian ideograms together with a picture that represented the meaning of the Kanji figure. In the recognition test, the children had to select the Asian ideogram that represented the correct meaning from among 4 ideograms, and rate their confidence in performance accuracy on a seven point scale. After they had received feedback, the children again expressed their confidence in performance accuracy on a thermometer scale. Finally, they had to give themselves credit points, taking into account the visually represented feedback. The complete session was repeated a week later with similar Kanji figures and the data from the 2 sessions were aggregated. In line with the researchers' expectations, self-evaluations were less accurate for the younger children, but they did make use of valid cues (i.e., item difficulty) to judge the accuracy of their performance. Contrary to expectations, both age groups were overconfident for incorrect responses but made adequate use of performance feedback to improve their self-evaluation. Yet, the older children were consistently less confident for incorrect outcomes than for correct responses, compared to the younger children. Also, the Kindergartners assigned themselves inappropriately high self-rewards for incorrect responses, despite the received negative feedback. The researchers concluded that their findings do not confirm previous research by Koriat and Ackermann that showed that children under 8 years cannot yet base self-evaluations on cues derived from the study experience. The fact that the younger children did not use the provided feedback to the same extent than the older children did when self-rewarding, was interpreted by the researchers as support for the self-protective bias hypothesis. Referring to the relevant literature, they point out that 6 year olds correctly understand the reward principle but that they apply it only for their peers and not for themselves. When reading the manuscript, I was wondering whether there was another explanation for the overconfidence of the younger age group and their inappropriate self-rewards.

The present experiment shows that both the younger and the older age group decoded the informational properties of the feedback correctly (i.e., they used the feedback to correct their self-evaluations). Regrettably, the researchers did not investigate whether there were developmental differences in the positive and negative emotions that were triggered by the feedback. Harter's (2006) developmental analysis revealed marked differences between children that are relevant to this study: below the age of 8, children show an inability to integrate attributes and emotions of *opposing valence* and they are inclined to overestimate their talents and skills. Between the ages of 8 and 11 children develop a representational system in which activities that elicit positive emotions are integrated with activities that trigger negative emotions, so that opposing valences may be considered simultaneously.

In light of these developmental differences, I wonder whether negative feedback had more impact on the older than on the younger children. I reckon that most of the older children in this study had already acquired 2 related school principles, namely (1) that the <u>outcome</u> of a learning experience should be used as the primary cue for self-reward, and (2) that negative feedback is consequential in nature. The point I am making is that the older children may have lowered their self-reward after receiving negative feedback, because the negative emotions triggered by the negative feedback intruded on their task specific enjoyment. In the same vein, I speculate that the younger children were still (partially) blind to the 2 school principles mentioned above, meaning that the negative feedback did not carry the same weight. It either did not elicit negative emotions or only very mild ones which did not intrude on their task specific enjoyment, because they could not deal with opposing valences. In other words, the positive emotions elicited during the learning episode lingered on in their information processing system and were used as the primary cue for self-reward (see also Fredrickson, 2004 who documented that positive emotions elicited during an activity give people a clear sense of purpose and a positive mind-set and are a strong antidote to negative emotions).

Raaymakers, Baars, Schaap, Paas, and Van Gog (2017) predicted that the feedback given after solving complex problems may change participants' appraisal of the mental effort they have invested in doing the set task. Such modification is problematic, because it raises doubts as to the value of 'mental effort' as an indicator of cognitive load. Raaymakers et al. (2017), designed an intervention to investigate the influence that feedback has on perceived mental effort. They asked participants with a broad age range to solve several 5 step problems on line. In the first experiment, participants had 1 min to solve each of the 5 complex, unfamiliar problems mentally. The participants had been randomly assigned to any of 3 feedback conditions: positive, negative and no feedback. The problems were difficult to self-assess which allowed the researchers to manipulate the FB. They predicted and found that students who received predominantly positive feedback immediately after finishing each of the five problems (e.g., PNPPP) would rate the invested effort lower than participants who predominantly received negative feedback (e.g., NPNNN) irrespective of their perceived level of learning. Although no significant difference in performance was found between the 3 feedback conditions, appraisals of mental effort decreased linearly across conditions from FB- to FB° and FB+. Experiment 2 was an attempt to replicate the results of experiment 1 and - in addition - the researchers investigated whether there was a timing effect. Dissimilar to experiment 1, there were 5 feedback conditions, namely (1) no feedback, (2) predominantly positive feedback with a reversal on the second task (PNPPP), (3) with a reversal on the fourth task (PPPNP), (4) predominantly negative feedback with a reversal on the second task (NPNNN) and (5) with a reversal on the fourth task (NNNPN). The valence of FB, irrespective of where in the sequence it was given, affected students' perception of the mental effort they had invested.

Raaymakers et al. (2017), wanted to replicate the results of experiment 1 with ecologically valid problems in a classroom setting. Higher education students were randomly assigned to the 3 feedback conditions. They worked in an on-line learning environment and started with an introductory video with a model example before solving the 5 biology problems with increasing complexity. After solving each problem, students had to self-assess their performance. Immediately afterwards feedback appeared on screen, according to the FB-condition they were in (FB = your answer has been registered; FB− = 2 points lower than their outcome appraisal; FB+ = 2 points higher than their outcome appraisal). Finally they had to rate their invested effort on a scale that ranged from 0 to 5. Although the effort ratings were overall lower than in experiments 1 and 2, a main feedback condition effect was noted. Students in the FB + condition scored significantly lower on reported mental effort than students in the FB- condition, thus replicating the results of experiments 1 and 2.

Taken together, the 3 experiments show that the valence of feedback has a prominent impact on reported effort. This implies that external feedback provided by somebody in authority was used as a cue to modify one's recollection of the effort invested in the problem solving process. It is a pity that the students were not asked for a confidence rating prior to receiving the feedback. This would have allowed the researchers to establish whether or not the students' outcome appraisal and/or their judgement of learning were overruled by the positive and negative feedback. An hypothesis that could be tested in future research is that the effect of feedback on mental effort was mediated by the *affect* triggered by the feedback. Positive feedback may have created positive affect (joy, relief or contentment) and this overall positive affective experience may have temporarily embedded the learning experience in a positive network of associations, prompting students to scale down the amount of effort. The reverse effect may have been produced by negative feedback: FB- may have triggered irritation, frustration, and/or tension and these negative emotions may have acted as a cue to scale up one's effort rating. Kuhl (2000) illustrated that when affect is mounting, the students' perceptions of the situation changes as well as their effort and competence appraisal. I will return to this point shortly (i.e., in the section on neglected aspects of SR).

Maranges, Schmeichel and Baumeister (2017) argued that 2 cognitive states, namely CL and ego depletion undermine deliberate, controlled, and complex thinking, because these states deplete resources. However, the resources that are affected by these 2 states are not the same. They set up a direct comparison of CL and ego depletion in 3 experiments. In a first experiment, undergraduate female students had to read a dull and dense medical text out loud. The control group was then asked to immerse their hand in icy cold water while the CL group was asked to do the same, but perform a concurrent task as well (count backwards in increments of 3). The ego depletion group had to read the same text, but they had to make an effort to express interest and enthusiasm in what they read. The hypothesis was confirmed: both CL and ego depletion influenced students' capacity to tolerate pain, but in the opposite direction. Counting backwards during the test increased persistence of holding the hand in ice water relative to the control condition. Having to pretend a text was interesting resulted in quitting faster. The second experiment involved a visual memory task with adult participants. The CL group had to remember a 10 digit numbers string while looking at a set of negative, positive, and neutral pictures. The ego depleted group was asked to write 2 short essays in a fixed period of time, but they were given strict instructions not to use certain letters. Students in all conditions were asked to report on the positive (i.e., alert, inspired, determined, attentive, and active) and negative emotions (i.e., upset, hostile, ashamed, nervous, and afraid) they had experienced while viewing the pictures before they got the visual memory test. Results showed no condition effect on visual memory, but the reported emotions were affected differentially. The CL group reported less negative and more positive emotions than the 2 other groups and the reversed pattern was found in the ego-depletion group. These results were confirmed in a 3rd experiment with a lexical judgement task and seem to be quite robust. The researchers explained that students in the depletion group had used self-control to follow the complicated instructions. This had temporarily depleted their self-regulatory resources, which made it more difficult for them to ignore the aversive stimulus in experiment 1 (i.e., quitting faster), and override the negative feelings during the visual memory and lexical judgement tasks. Dissimilarly, students in the CL group had to perform a concurrent task, which increased the processing demands in working memory. This manipulation reduced the impact of the aversive stimulus in experiment 1 (i.e., it increased persistence), and decreased the negative feelings and increased the positive feelings in experiments 2 and 3.

I certainly find that Maranges et al. (2017), made a valuable contribution to the field. They introduced ego depletion in the educational psychology literature and I am sure that future studies will underscore the importance of this construct for classroom research. However, my interpretation of the findings is slightly different. Requesting students to focus on a concurrent task while they hold their hand in ice cold water is in fact prompting them to use an emotion regulation strategy. As explained by Boekaerts and Pekrun (2016) it is often extremely difficult for students to ignore intrusive thoughts and upsetting information in the learning environment, because the monitoring system keeps focusing on the disturbing stimuli. A way to get around negative monitoring is to use a coping or emotion regulation strategy, such as 'reappraisal', 'mental distraction' and 'mental disengagement'. These and similar coping strategies redirect the focus of the students' attention (i.e., a shift in monitoring strategy). It is important to note in this respect that a distracting mental task may free WM from worry and tension associated with the primary task (e.g., in experiment 1), thus creating more rather than less room for cognitive processing. In my judgment, it needs to be tested whether the observed effects were due to increased processing demands (i.e., creating CL) or due to the use of an emotional regulation strategy called "mental distraction. More precisely, were the students in the CL group more alert, inspired, attentive and active and less hostile and upset relative to the students in the other groups due to the induced CL or because they used an effective SR strategy?

Taken together, the 6 papers provide examples of how different research groups conceptualize and measure CL and how they try to link CL to SR. Each contribution shows a different approach and each approach has its strong and weak points. However, there are some general issues that beg for greater clarification and elaboration if we want to gain better insight into the relation between CL and SR, namely (1) is there consensus on the meaning and measurement of CL? (2) What methodologies can be used to capture CL in real time and link it to strategy use, (3) what key aspects of SR remained underexposed in the different manuscripts, and (4) what cues do students pick up to adapt to increased and decreased processing demands. I will address each of these issues in turn.

2. Is there consensus on the meaning and measurement of cognitive load?

Let me inform the reader that I am not an expert on CL. When reading the articles, I looked for the authors' definition and operationalization of CL, but not all research groups provided a working definition. They all seemed to assume that overload leads to reduced learning and that this is induced when the total amount of cognitive processing that is needed exceeds the students' maximum cognitive processing capacity. Yet, none of the research groups conceptualized and operationalized CL in such a way that it was transparent for the reader how much processing load that performing the respective tasks had imposed on the student's cognitive system. In the absence of explicit information about the expected cognitive capacity demands of the tasks and without information about the students' working memory capacity (WMC), it is impossible to establish whether or not the respective tasks imposed an optimal or suboptimal CL on the participants. We also do not know whether the students' cognitive processing capacity was exceeded in any of the experiments reported in the Special Issue. In fact, I was struck by the lack of conceptual clarity which extends to the definitions and measurement instruments.

Each research group selected an instrument from the available list of instruments that measures whether or not CL occurred. Surprisingly, most researchers picked this instrument without specifying the specific components of CL that they intended to measure (e.g., overall cognitive load, germane load, intrinsic load, extraneous load). I have serious doubts that the various self-report measures used by the different research groups to measure aspects of CL did indeed capture CL in a valid and reliable way (see also Moreno, 2010). For example, Schleinschok et al. (2017), used a 2 item questionnaire to assess CL in terms of the students' appraisal of the *difficulty level* of the learning tasks. Granted, self-reports of perceived level of difficulty of a task gives us an indication of how the students perceive and interpret the task, but this appraisal is only a naïve way of assessing the complexity and difficulty level of the task in terms of the students' confidence judgement that they can answer test questions on the paragraph. Schleinschock et al. (2017), reported that the correlation between the students' difficulty appraisal and their judgement of learning (Jol) was quite strong, indicating that these two appraisals are measuring quasi the same thing. Hence, we may doubt whether the criteria that the students used to rate the difficulty level of the respective tasks were different from the criteria they used to assess their confidence that they understood what was written in the respective paragraphs. My distrust in using perceived difficulty level as a measure of CL stems from a finding reported by Crombach, Boekaerts, and Voeten (2003). They tested the internal structure of Boekaerts' appraisal model in the math domain using Lisrel and concluded that 2 of the students' task appraisals, namely their difficulty appraisal and their subjective competence appraisal were not empirically distinguishable.

I was more enchanted by the way Van Loon et al. (2017), operationalized CL. They used the objective difficulty level of the Kanji problems to determine each task's intrinsic CL (i.e., the difficulty index of a Kanji is equal to the percentage of correct answers for a particular item across a separate sample of participants in the same age range, corrected for the probability of guessing). My position is that researchers who want to know what the processing demands are of the tasks they selected for their studies need to determine task complexity in terms of pre-established categories of complexity before the experiment starts. For example, several researchers have defined task complexity in terms of the multidimensional structure of a problem or task, describing a problem in terms of the multiple paths to its solution and the uncertainty of possible outcomes (see Garden et al., 2006 who determined the complexity index of all math tasks used in the TIMMS project; Musso, 2016).

Glogger-Frey et al., 2017 explicitly mentioned that they measured *extraneous* load. Using, Leppink, Paas, Van der Vleuten, Van Gog, and Merrienboer (2013) five item scale, they examined whether the students who worked in the invention condition experienced more extraneous load than students in the guided condition, and whether the perceived extraneous CL had impaired their performance on a transfer task. Interestingly, this research group took account of individual differences in beliefs about their math and science performance and reported that these beliefs interacted with CL. Students who believed that their self-perceived skills in math and science were adequate were less affected by extraneous CL than students who judged their competence to be low.

Sidi et al. (2017), used time pressure to induce CL and compared the performance of students who worked on screen and on paper. It is not clear whether TP affects all students in the same way. Hence, it needs to be verified whether some students were cognitively overloaded and others still functioned at an acceptable level of CL before drawing firm conclusions. Finally, Maranges et al. (2017), selected a concurrent task (i.e., counting backwards and remembering a 10 digits span) to induce CL. It is difficult to judge how the concurrent task interacted with the students' visual encoding of the positive, neutral and negative pictures. The task may have triggered emotion regulation strategies in addition to an increase in processing demands.

Raaymakers et al. (2017), used Paas's one item mental effort scale to assess CL. It is not clear to me how untrained participants recruited on-line rated the amount of effort invested in these unfamiliar tasks? Should simply doing the task be rated as 'no effort at all' or should it be rated as 'average effort'? In line with the comments levelled at the measurement of difficulty level, I have doubts that the mental effort scale captured CL in a reliable and valid way. Apart from the fact that the psychometric properties of a one-item mental effort scale cannot be determined, it should not be assumed that all participants can accurately rate the mental effort they invested in a task. Their naïve judgements of the mental effort they invested may not even align with what actually happened during problem solving. At the very least, participants need to be trained to use explicit criteria to accurately rate how much effort they invested to meet the demands of the imposed task. I would recommend researchers on CL to determine the processing demands of the tasks they set in terms of their complexity and difficulty level. By arranging the tasks in order of increasing complexity and/or difficulty level, they can then determine at what point in the sequence the students' WMC is exceeded (see my discussion on multiple time points).

At this point in the discussion I would like to remark that there are wide individual differences in WMC and attentional resources and by implication in the effect that processing additional information and performing concurrent tasks may have on performance. Without an indication of the students' WMC, the experimenter has no inkling of how much processing capacity is used. Admittedly, some highly complex tasks and instructional procedures may create cognitive overload in all students but we need fine grained information on whether or not the CL induced by the target task, the extra task, and/or the instructional procedures exceeded the WMC of *all* the students or only of students with low WMC. In the latter case, the performance of students with mid to high WMC, as well as their score on various process measures, would not be expected to be different from that of the students in the control group. Although assessment of individual differences in WMC and attentional resources is possible, this type of objective measurement is absent from most educational psychology research. A notable exception is the work by Musso (2016). She used a computerized task, called AOSPAN (Unsworth, Heitz, Schroch, & Engle, 2005) to measure WMC. This is a valid and reliable indicator of WMC and has been used in a variety of research areas. Using this instrument, Musso (2016) found that individual differences in WMC influenced math problem solving performance directly. She also reported that positive appraisals - more specifically a positive subjective competence appraisal, as well as positive emotions experienced during the task - moderated the effect of cognitive processing capacity on math performance.

3. What methodologies can be used to capture CL in real time and link it to strategy use?

As mentioned previously, most of the studies presented in this Special Issue are limited to self-reports that estimate overall CL. Although I do not find fault with the use of self-reports when the object of investigation is a non-observable state or process, I think that self-reports should always be flanked by other measurement tools, such as tracks left behind while working, performance indicators, observations by trained professionals, in-depth interviews, and stimulated recall video sessions. Retroactive appraisals cannot accurately register what has actually happened during the learning process, because students cannot recall the what, how, when and why of the learning process in detail. This criticism can also be levelled at the new measurement instruments that have become available recently (Leppink et al., 2013) to supplement overall estimates of CL. These self-reports measure different types of CL, namely the degree of mental effort invested (1) in the complexity of this activity (intrinsic CL), (2) in enhancing my knowledge and understanding (germane load) and (3) in unclear and ineffective explanations and instructions (extraneous load). Granted, these differential questions for the three types of load definitely upgraded the overall measures of CL and will help to disambiguate the theoretical distinctions. Yet from my experience with administering questionnaires to students, I doubt whether all students can differentiate between these theoretical distinctions. Even if they comprehend all the items, their answers leave us in the dark as to the underlying processes.

This brings me to an essential aspect of measuring CL, namely the fact that it is *contextualized* and consequently that it does not suffice to ask retrospectively whether CL occurred. It may have occurred several times during the learning episode, which calls for a design with multiple time points. For example, students who solved a series of problems may have experienced several increases in task demands at different points in the sequence. At one time point they may have become more alert and met the challenge with a shift of monitoring strategy. At another time point, they may have avoided distractions. And, at yet another time point they may have downscaled their current goal or even given up on the task altogether. Likewise, a felt decrease in task demands may have resulted in coasting behaviour, reflected in being more relaxed, monitor less accurately, choosing surface level learning strategies, and adopting a different goal. Clearly, these adaptations may alter the learners' cognitive and affective states, including their awareness of CL, which makes it difficult to report it retrospectively.

A major challenge for CL research is to come up with measurement instruments that can track shifts in perceived task demands in real-time and provide detailed information about the unfolding cognitive processes and SR strategies that students use to counteract increased or decreased processing demands. In the studies reported in this Special Issue, the researchers treat the self-regulation process as one holistic unit and ignore the sequential and temporal characteristics of the learning experience. For example, in the study by Glogger-Frey et al., 2017 the experiment consisted of a sequence of different learning activities (i.e., questionnaires, 1st training session, questionnaires, tests, 2nd training session, direct instruction, solve transfer problems). This implies that the students may have experienced different challenges and threats during the experiment (e.g., experienced as low confidence, awareness of gaps in knowledge, frustration because of task complexity and confusing instructions). It is evident that these changes in the learners' cognitive and affective states influence each other over time. It is therefore advisable that in future research on the interface between CL and SR the sequential and temporal aspects of the students' involvement with the learning activity are taken into account. This would allow the researchers to examine where exactly in the learning sequence cognitive (over) load occurred, its duration, and the rate at which it occurred *in close connection with* the SR processes that the students used to counteract it. For further discussion of these types of analyses see Azevedo, 2015; Greene and Azevedo 2010; Molenaar, 2014; Molenaar & Chiu, 2014; Molenaar & Järvelä, 2014; Winne, 2014.

Apart from analyzing interaction data from traditional computer tracking of the learning process, CL researchers may also want to familiarize themselves with interaction data from working memory load-related measures, such as EEG alpha frequency band power, eye-movement data, pupil diameter, multi-touch interactions, heart rate, skin conductance, and neurophysiological data recorded from the scalp (for early introductions of psychophysiological measures into the working load literature, see Paas, Van Merriënboer & Adam, 1994; Paas, Tuovinen, Tabbers, & van Gerven, 2003). Recently, Gerjets and colleagues, working at the Leibnitz Institute für Wissensmedien in Tübingen pointed out that learning applications that run on devices with an interactive screen (e.g., multi-touch tables) have built-in sensors that register intuitive gesture-based interactions during the learning process, such as touches, swipes, and pinches to zoom in and out, which can be used to determine cognitive workload. Gerjets and colleagues collected interaction patterns of primary school children while they solved math problems on a multi-touch table and then used machine learning methods on these high-resolution interactions (see Mock et al., 2016). The results showed that touch screen interaction patterns can be used to predict high levels of cognitive workload induced by tasks of varying difficulty levels, even without knowing anything about the learners' speed and accuracy of task performance. Interestingly, increased cognitive workload manifested itself differently for different students. Additionally, Gerjets and his team demonstrated that EEG alpha frequency band power and pupil dilation are also sensitive to register increased processing load on executive functions, such as updating and inhibition (Scharinger, Soutschek, Schubert, & Gerjets, 2015). The most promising avenue to monitor

CL seems to rely on neurophysiological data (i.e., data from skin electrodes on the scalp). Gerjets (2017) predicted that in the near future, high resolution touch sensor interaction data may be used for the automatic assessment of cognitive and affective learner states and that methods of graded workload detection will become available to help develop optimal math learning environments by means of real-time EEG workload adaptation of the learning tasks. It is certainly promising for research groups who investigate CL that their studies need not be restricted to measuring CL with self-report measures. Multiple methodologies for studying the effect of CL on different process and outcome variables in real time have been brought to the research table.

4. Key aspects of SR that remained underexposed in the different manuscripts

I hope it is clear from my discussion of the various papers, that I regret that affect still takes a back seat in research on CL and SRL. The same holds for goals and motivation regulation strategies. While reading the various articles I continually asked myself why so little attention had been given to the 3rd layer of self-regulated learning as described by Boekaerts more than 20 years ago (Boekaerts, 1997; 1999). She described SRL as a three layered process (see Fig. 1). The core or inner layer of SRL refers to the learning process per sé (self-regulated learning), or more concretely to the way students process information and hence to the selection and organization of the cognitive strategies that make up the learning process. The middle layer refers to the regulation of the learning process (self-regulated learning), meaning to the choice of metacognitive strategies that direct, monitor, and control the learning process. The outer layer refers to the self (self-regulated learning), more specifically to the way students protect the learning process from competing action tendencies. Their choice of goal(s) based on their beliefs about learning and social interactions in the classroom and the use of motivation regulation strategies are central to this layer.

Boekaerts and Corno (2005) argued that self-set goals are the driving force for choosing cognitive and metacognitive strategies that direct the learning process and for selecting motivation and emotion regulation strategies that support this choice, especially when obstacles and distractions make it difficult to protect the learning goal from competing action tendencies. Using this model as a framework for classifying the papers in the Special Issue, I conclude that 5 of the 6 papers deal with the middle and inner layer of the model (regulated learning) and that only the paper by Maranges et al. (2017), considers aspects of the outer layer.

It is important to realize in this respect that most of the tasks that the students had to do in the respective studies were set in controlled experimental settings. The students had to work on *imposed* goals, which they may or may not have perceived as valuable and instrumental to increase their own resources (Ryan & Deci, 2009). A vast body of literature documents that the beliefs students hold about learning in and out of school — especially their value and competence beliefs - are conditional for the use of cognitive (Entwistle, 1998), metacognitive (Efklides, 2011) and motivation regulation strategies (e.g.,Wolters, 2003). In light of these findings, I would recommend that CL researchers select ecologically valid task and measure students' situation-specific value and competence appraisals (see Boekaerts, 2002), as well as their *self-set goals* prior to the problem solving process, and again after task completion. If multiple measurement points are used, it would allow researchers to detect shifts in beliefs during the experiment, and to examine how, when, and why these shifts have reshaped the students' goal(s) and the SR strategies derived from these goals. This would answer relevant questions, such as: why did the explanations that students gave in the guided condition in Glogger-Frey et al., 2017 's study deteriorate? Did they perceive the 2nd training session as less valuable? Did they feel overconfident? Did they experience negative affect?

5. What cues do students pick up to help them adapt to increased and decreased processing demands?

Sidi et al., pointed out, that students are aware that different learning tasks and learning contexts have different purposes. In other words, they do not choose their goals randomly. Rather, they look for *cues* that may inform them what is expected, possible given the context, and legitimate. If such cues are detected, students interpret them and make choices based on this interpretation (e.g., select strategies from the repertoire of cognitive and metacognitive

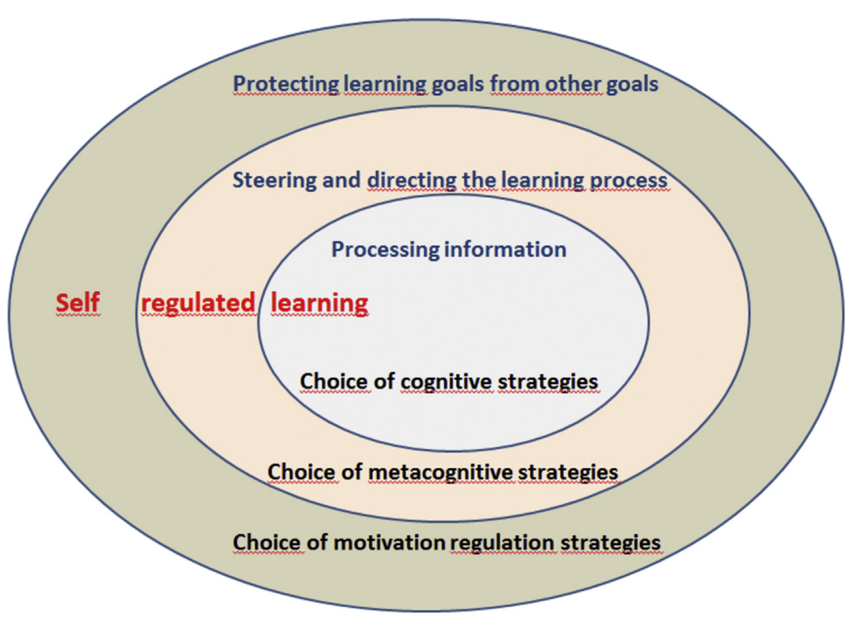

Fig. 1. Three layered model of SRL, depicting the functions of diverse self-regulation strategies.

strategies to start the task, and also choose motivation regulation strategies to complete the task). I enthusiastically embrace the cue-utilisation framework that the editors of the Special Issue proposed for setting up joint research (De Bruin & Merrienboer, 2017). A cue-driven model can identify the cues that students of different age groups *actually* use to detect an increase or decrease in processing load. I am confident that using such a model will increase our understanding of the cues that trigger students' choices of cognitive and metacognitive strategies and the cues that prompt them to regulate their motivation in order to enact this choice. To get a comprehensive picture of the cues that work for students, it is desirable that theory driven cue identification is supplemented by a bottom-up procedure: i.e., asking students to report on their own cues, because what works for the average student may not work for all students. I speculate that the latter procedure will reveal that many environmental cues, such as feedback, trigger positive and negative emotions that interfere with the students' choice of strategies. Evidence from cognitive psychology and neuroscience shows that environmental cues may trigger goals outside of awareness, which then run automatically to attain desired outcomes (Custers & Aarts, 2010). These findings bring up the question: Do goals that are primed by environmental cues compete for processing resources with goals that are intentionally pursued?

6. Some final thoughts

In conclusion, I express my thanks to the editors of this Special Issue for bringing together researchers from 2 related domains of research and for giving me the chance to write this commentary. I also thank the different research teams for a stimulating set of papers. Reading these papers allowed me to tie some loose ends in my understanding of why students do or do not self-regulate their learning. Although much remains to be accomplished in this important area of research, the different contributors showcased excellent research that illustrates the complexity of CL as a construct. This is reflected in the absence of a clear conceptualization and in the divergent ways that it is measured across the papers. It is important that researchers agree on the meaning and measurement of CL and on how it is related to SRL. Explanatory models need to be developed, detailing the psychological processes underlying students' understanding and coping with cognitive load. These models must also account for the fact that in most modern classrooms students learn together, be it in dyads or in groups, and that the resulting interaction patterns between students may alleviate or aggravate CL. Together, the manuscripts provided new perspectives that can advance the studies of CL ànd SR, and more importantly, on the type of bridge(s) that could connect CL research to self-regulated learning. I hope that my comments and suggestions are useful to support this important mission.

References

Azevedo, R. (2015). Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. *Educational Psychologist, 50*(1), 84.

Boekaerts, M. (1997). Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. *Learning and Instruction, 7*(2), 161–186.

Boekaerts, M. (1999). Self-regulated learning: Where we are today. *International Journal of Educational Research, 31*(6), 445–457.

Boekaerts, M. (2002). The on-line motivation questionnaire: A self-report instrument to assess students' context sensitivity. In *Advances in Motivation and Achievement* (Vol. 12, pp. 77–120). New York: Elsevier.

Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied psychology: An International Review, 54*(2), 199–231.

Boekaerts, M., & Pekrun, R. (2016). Emotions and emotion regulation in academic settings. In L. Corno, & E. M. Anderman (Eds.), *Handbook of educational psychology* (pp. 76–90). New York: Routledge.

Crombach, M. J., Boekaerts, M., & Voeten, M. J. M. (2003). Online measurement of appraisals of students faced with curricular tasks. *Educational and Psychological Measurement, 63*(1), 96–111.

Custers, R., & Aarts, H. (2010). The unconscious will: How the pursuit of goals operate outside of conscious awareness. *Science, 329*(5987), 47–50.

De Bruin, A. B., & van Merriënboer, J. J. (2017). Bridging Cognitive Load and Self-regulated Learning Research: A complementary approach to contemporary issues in educational research. *Learning and Instruction, 51*, 1–9.

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: MASRL model. *Educational Psychologist, 46*(1), 6–25.

Entwistle, N. (1998). Approaches to learning and forms of understanding. In B. Dart, & G. Boulton-Lewis (Eds.), *Teaching and learning in higher education* (pp. 72–101). Melbourne: Australian Council for Educational Research.

Fredrickson, B. L. (2004). The broaden-and-build theory of positive emotions. *Philosophical Transactions Royal Society London B, 359*, 1367–1377.

Garden, R. A., Lie, S., Robitaille, D. F., Angell, C., Martin, M. D., Mullis, I. V., et al. (2006). *TIMMS advanced 2008 assessment framework. International association for the evaluation of educational achievement*. Boston USA: TIMMS & PIRLS International Study Center.

Gerjets, P. (2017). Learning and problem solving with hypermedia in the twenty-first Century: From hypertext to multiple web sources and multimodal adaptivity. In S. Schwan, & U. Cress (Eds.), *The psychology of digital learning* (pp. 61–88).

Glogger-Frey, I., Gaus, K., & Renkl, A. (2017). Learning from Direct Instruction: Best prepared by several self-regulated or guided invention activities? *Learning and Instruction, 51*, 26–35.

Greene, J. A., & Azevedo, R. (2010). The measurement of learners' self-regulatory cognitive and metacognitive processes while using computer-based learning environments. *Educational Psychologist, 45*, 203–209.

Harter, S. (2006). The self. In W. Damen, & R. M. Lerner (Eds.), *Handbook of child psychology* (Vol. 3, pp. 505–570). Hoboken NJ: Wiley.

Kuhl, J. (2000). A functional-design approach to motivation and self-regulation: The dynamics of personality systems and interactions. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 111–169). San Diego: Academic Press.

Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T., & Merrienboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods, 45*, 1058–1072.

Maranges, H., Schmeichel, B., & Baumeister, R. (2017). Comparing Cognitive Load and Self-regulatory Depletion: Effects on Emotion and Cognition. *Learning and Instruction, 51*, 74–84.

Mock, P., Gerjets, P., Tibus, M., Trautwein, U., Möller, K., & Rosenstiel, W. (2016). Using Touchscreen Interaction data to predict cognitive workload. In Y. I. Nakano, E. André, T. Nishida, L. P. Morency, C. Busso, & C. Pelachaud (Eds.), *Proceedings of the 18th ACM international conference on multimodal interaction* (pp. 349–356). New York, NY: ACM.

Molenaar, I. (2014). Advances in temporal analysis in learning and instruction. *Frontline Learning Research, 2*(4).

Molenaar, I., & Chiu, M. M. (2014). Dissecting sequences of regulation and cognition: Statistical discourse analysis of primary school children's collaborative learning. *Metacognition and Learning, 9*(2), 137–160.

Molenaar, I., & Järvelä, S. (2014). Sequential and temporal characteristics of Self and Social regulated learning. *Metacognition and Learning, 9*, 75–85.

Moreno, R. (2010). Cognitive load theory: More food for thought. *Instruction Science, 38*, 135–141.

Musso, M. F. (2016). *Understanding the underpinnings of academic performance*. Belgium: Doctoral Disseration defended at the Katholieke Universiteit Leuven.

Paas, F., Tuovinen, J., Tabbers, H., & van Gerven, P. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63–71.

Paas, F., van Merriënboer, J. J. G., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills, 79*, 419–430.

Raaymakers, S. F., Baars, M., Schaap, L., Paas, F., & Van Gog, T. (2017). Effects of Performance Feedback Valence on Perceptions of Invested Mental Effort. *Learning and Instruction, 51*, 36–46.

Ryan, R. M., & Deci, E. L. (2009). Promoting self-determined school Engagement: Motivation, learning, and well-being. In K. R. Wentzel, & A. Wigfield (Eds.), *Handbook on motivation at school* (pp. 171–198). New York: Routledge.

Scharinger, C., Soutschek, A., Schubert, T., & Gerjets, P. (2015). When flanker meets the n-back: What EEG and pupil dilation data reveal about the interplay between the two central-executive working memory functions inhibition and updating. *Psychophysiology, 52*, 1293–1304.

Schleinschok, K., Scheiter, K., & Eitel, A. (2017). Do drawing tasks improve monitoring and control during learning from texts? *Learning and Instruction, 51*, 10–25.

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.

Unsworth, N., Heitz, R. P., Schroch, J. C., & Engle, R. W. (2005). An automated version of the operatio6n span task. *Behavior Research Methods, 37*(3), 498–505.

Van Loon, M. H., Destan, N., & Spiess, M. (2017). Developmental Progression in Performance Evaluations: Effects of children's cue Utilization and Self-Protection. *Learning and Instruction, 51*, 47–60.

Winne, P. H. (2014). Issues in researching self-regulated learning as patterns of events. *Metacognition and Learning, 9*, 229–237.

Wolters, C. (2003). Regulation of motivation: Evaluating an underemphasized aspect of self-regulated learning. *Educational Psychologist, 38*, 189–205.