Instr Sci (2015) 43:93–114
DOI 10.1007/s11251-014-9328-3

Measuring cognitive load with subjective rating scales during problem solving: differences between immediate and delayed ratings

Annett Schmeck · Maria Opfermann · Tamara van Gog · Fred Paas · Detlev Leutner

Received: 28 November 2013 / Accepted: 21 July 2014 / Published online: 21 August 2014

Abstract Subjective cognitive load (CL) rating scales are widely used in educational research. However, there are still some open questions regarding the point of time at which such scales should be applied. Whereas some studies apply rating scales directly after each step or task and use an average of these ratings, others assess CL only once after the whole learning or problem-solving phase. To investigate if these two approaches are comparable indicators of experienced CL, two experiments were conducted, in which 168 and 107 teacher education university students, respectively, worked through a sequence of six problems. CL was assessed by means of subjective ratings of mental effort and perceived task difficulty after each problem and after the whole process. Results showed that the delayed ratings of both effort and difficulty were significantly higher than the average of the six ratings made during problem solving. In addition, the problems we assumed to be of higher complexity seemed to be the best predictors for the delayed ratings. Interestingly, for ratings of affective variables, such as interest and motivation, the delayed rating did not differ from the average of immediate ratings.

Keywords Cognitive load · Problem solving · Mental effort · Task difficulty · Measurement

Introduction

In recent years, cognitive load theory (CLT; Sweller 2010; Sweller et al. 1998; Van Merriënboer and Sweller 2005), with its central assumption that learning is a function of

A. Schmeck (✉) · M. Opfermann · D. Leutner
Department of Instructional Psychology, Faculty for Educational Sciences, University of Duisburg-Essen, 45117 Essen, Germany
e-mail: annett.schmeck@uni-due.de; annett.schmeck@uni-duisburg-essen.de

T. van Gog · F. Paas
Capaciteitsgroep Psychologie, Faculteit der Sociale Wetenschappen, Erasmus University Rotterdam, Burgemeester Oudlaan 50, T-gebouw, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

available cognitive resources, has become increasingly important in educational research. Accordingly there is a large body of empirical studies inspired by CLT that support its implications for instructional design (for overviews see de Jong 2010; Paas et al. 2003a, b; Plass et al. 2010; Sweller et al. 2011). Measuring cognitive load (CL) plays a central role in instructional design research. However, although subjective rating scales are widely used to measure CL, little is known about when to apply these rating scales and how this affects the ratings. The two studies reported in this paper intend to shed some light on the question of whether subjective load ratings are influenced by the point of measurement.

Cognitive load and instructional design

According to the central tenet of CLT (Sweller 2010; Sweller et al. 1998; Van Merriënboer and Sweller 2005), working memory capacity is limited, and the aims of any instructional design should be to reduce unnecessary working memory load to free capacity for learning-related activities, that is, schema construction. Working memory load was initially divided into intrinsic cognitive load and extraneous cognitive load (Chandler and Sweller 1991). Intrinsic cognitive load refers to the load imposed by cognitive processes evoked by task-inherent characteristics, or in short: the load that relates to task complexity. How complex a task is perceived to be, in turn, depends on prior knowledge as well as on element interactivity (which means the number of elements that have to be processed simultaneously and in relation to each other when performing a task). On the other hand, extraneous cognitive load is imposed by cognitive processes evoked by the instructional design of a learning task that do not contribute to learning. For instance, when a task contains many seductive but task-irrelevant details (Park et al. 2011), cognitive resources are "wasted" by paying attention to those details and are thus not available for learning-related activities. Imagine, for example, presentation slides in a lecture, where each sentence has a different color or where headlines are blinking and moving. This may look nice (and is thus called seductive), but when it does not directly contribute to learning or understanding, it should not be used. The same applies to relevant information sources that are to be found far from each other, such as textual explanations and according graphics on different pages in a textbook. In this case, they would be causing a so-called split-attention effect (Chandler and Sweller 1992), that is, cognitive capacity would again be "wasted" merely for visual search processes, while it could actually be better used for comprehension of the materials.

Because of the aforementioned limitations of working memory and because both types of load are assumed to be additive, the goal of instructional designers should thus be to keep extraneous cognitive load as low as possible, so that sufficient working memory capacity is available for processes that foster learning. The load that is caused by such learning-relevant processes refers to the concept of germane cognitive load—this third type of working memory load was introduced later in the theory (Sweller et al. 1998; see also Sweller 2010). In more recent work, however, the term "germane cognitive load" is seen rather critical. For instance, Sweller et al. (2011) argue that "it is probably inappropriate to use this term" because "Unlike intrinsic and extraneous cognitive load […], germane cognitive load is not imposed by the learning materials" (p. 57). Rather, it is suggested to talk of *germane resources* that should be high enough to deal with the intrinsic cognitive load caused by the contents to be learned and given that the extraneous load caused by the (inappropriate) instructional design is not too high.

Measurement of cognitive load

The question of how to assess cognitive load most validly and reliably is still leading to heated discussions among researchers. In a recent overview on cognitive load measurement, Sweller et al. (2011) distinguish between indirect measures, subjective measures, secondary task measures, and physiological measures.

Indirect measures, such as computational methods, performance, or error profiles between problems can serve as indicators for the load experienced, but are no direct reflection of the actual cognitive load. For instance, Ayres (2006) found that error rates during mathematical problem solving increased when sophisticated decision-making came into play, that is, when many variables needed to be taken into account at once.

Measuring cognitive load by means of a secondary task, that is with dual task methodologies, is one of the frequently used methods (see for instance Brünken et al. 2003, 2004; Chandler and Sweller 1996; Park et al. 2011; Sweller, 1988). As a main advantage, Sweller et al. (2011) mention the possibility to receive continuous measures of cognitive load during task performance, which gives insight into the fluctuation of cognitive load over time. However, this advantage might be outweighed by the more complex experimental design that this method needs (Sweller et al. 2011), and its potential intrusiveness when applied during learning or problem solving tasks.

The latter disadvantage also applies to the third possibility, that is, the use of physiological indicators to measure cognitive load, such as heart rate (Paas and Van Merriënboer 1994), eye movements (Van Gerven et al. 2004; Van Gog and Jarodzka 2013) or brain activities (Antonenko and Niederhauser 2010; Antonenko et al. 2010; Paas et al. 2008). Altogether, Sweller et al. (2011) see physiological measures and in particular eye tracking results as "promising, but it is still too early to determine whether the current research emphasis will result in solid results" (p. 81).

The method that will be used in this study, however, and which also seems to be the preferred method in much of the current research is to assess cognitive load by means of subjective rating scales. In this regard, one of the measures that have been used very early in research are mental effort rating scales (see Paas 1992; Paas et al. 2003a, b; Van Gog and Paas 2008), which typically ask learners to rate the amount of mental effort they invested in completing a task on a 7- or 9-point Likert scale, ranging from "very, very low mental effort" to "very, very high mental effort". Another widely used subjective measure of cognitive load are ratings of perceived task difficulty (Bratfisch et al. 1972; Kalyuga et al. 1999; Marcus et al. 1996; Paas et al. 2003a, b), which typically ask learners to rate their perceived difficulty of a task on a 7- or 9-point Likert scale, ranging from "very, very easy" to "very, very difficult". Note that while perceived task difficulty and perceived mental effort may correlate, they are different constructs (see Van Gog and Paas 2008, for a discussion).

On the one hand, these subjective measures have been criticized for assessing cognitive load with only single items (e.g., Brünken et al. 2003). However, in many of the early studies, the item was repeatedly applied. Moreover, these studies showed the effectiveness of the rating scale by showing that the variation in learners' cognitive load ratings depended on variations in task complexity or instructional design (Ayres 2006; Paas et al. 1994; for overviews see Paas et al. 2003a, b; Van Gog and Paas 2008). In this regard, Sweller et al. (2011) conclude that "the simple subjective rating scale, regardless of the wording used (mental effort or difficulty), has, perhaps surprisingly, been shown to be the most sensitive measure available to differentiate the cognitive load imposed by different instructional procedures (p. 74).

However, there are still some open questions regarding the use of subjective rating scales for cognitive load measurement, especially regarding the point of time when such scales should be applied (Van Gog et al. 2012). On the one hand, many studies assessed cognitive load directly after each problem solving step or task, and used the average of these cognitive load ratings as an indicator of the overall cognitive load imposed by the learning task(s) (e.g., Opfermann 2008; Paas 1992; Paas and van Merriënboer 1994; Tabbers et al. 2004; Van Gog et al. 2006). On the other hand, there is a considerable amount of research that has assessed cognitive load with a single delayed measure; that is, cognitive load was measured only once after the whole learning/problem solving phase (e.g., Kalyuga et al. 2001; Kühl et al. 2011; Leutner et al. 2009; Schwamborn et al. 2011; Seufert and Brünken 2006).

Up to now, it seems that both immediate and delayed techniques have their advantages and disadvantages (Kuusela and Paul 2000; Taylor and Dionne 2000; Van Gog et al. 2012). Using one single delayed rating at the end of a learning or problem solving phase is quite economic; however, it is also rather difficult to determine what exactly this rating indicates. That is, it is unclear whether participants estimate their cognitive load as an average over all tasks, the last tasks they worked on (i.e., a recency effect), the most complex tasks they worked on, or any combination of those possibilities. In comparison, measures of cognitive load that are directly applied after each problem solving step or task, interact more with the problem solving step or task, as learners have to retrospect only on the task they just finished and that is probably still (partly) activated in working memory (*cf.* Van Gog et al. 2012). Sometimes, it seems as if immediate and delayed measures are often used arbitrarily and as if the average of ratings given during learning is the same as one single delayed rating. The question, however, is whether these two methods are comparable indicators of the actual load experienced. In other words, are there differences between the average of immediate scores and the delayed, overall score of cognitive load or not? So far, only a series of four experiments developed by Van Gog et al. (2012) tried to shed some light on this question. Results from their first experiment (between-subjects design) and their second experiment (within-subjects design), in which they used different arrangements of simple and complex problem solving tasks, showed that a single mental effort rating after a series of tasks (i.e. one delayed rating) resulted in a higher mental effort score than the average of mental effort ratings provided immediately after each task (i.e. immediate ratings). A similar result was shown in the third experiment, however, only with a series of complex tasks, not with simple tasks. Results of the fourth experiment revealed that knowing beforehand that a mental effort rating will be required after completing all tasks results in lower scores, but average delayed ratings per task still differed from a single delayed rating. In short, the results of Van Gog et al. (2012) provide first evidence that there are differences between the average of immediate scores and the delayed overall score of mental effort ratings. The question, however, what might cause these differences and whether they also apply to other widely used subjective measures of cognitive load, such as perceived difficulty, still remains unanswered. Furthermore, the experiments in the Van Gog et al. (2012) study focused on the cognitive load measures and on the differences between immediate and delayed ratings. At this point, it is unclear to what extent immediate and delayed ratings actually relate to performance measures and learning outcomes. For instance, Van Gog et al. (2012, p. 838) state that "it would be interesting to investigate in future studies what the relationship is between multiple ratings or single ratings and learning outcomes", and this point was taken up in our studies as well.

Another interesting aspect that was not investigated in the Van Gog et al. (2012) study is the question of why the delayed cognitive load ratings seem to be consistently higher than

the average of immediate ratings given after each task or problem. It might be possible that learners primarily "remember" the most difficult problems when they are asked to give a retrospective overall cognitive load rating. However, in their study, only very simple and very complex problems were used. By using problems at varying degrees of complexity, this question could be addressed, as this would allow for conducting hierarchical regressions to see which of the individual ratings given during learning best predict the retrospective overall rating. Finally, by not only assessing the two subjective cognitive load ratings, but also other subjective ratings, for instance, of affective variables (interest and motivation) after each task and after the entire series of tasks, it can be investigated whether the higher delayed rating is specifically associated with cognitive load measures or whether it is a general issue. Establishing this is important for finding the underlying mechanisms of the effect.

To sum up, the present studies do not only aim to replicate the findings regarding mental effort ratings firstly presented by Van Gog et al. (2012), but also to extend their research by including the other widely used cognitive load rating of perceived difficulty, by relating the cognitive load measures to performance measures in order to find out what might cause the difference between the immediate and delayed ratings, and by including not only subjective measures of cognitive, but also of affective variables in order to find out whether the effect is specific to the load ratings.

Research questions

The main question addressed in the present study is whether there are differences between the average of multiple ratings, that is, ratings given directly after each task, and a delayed rating of cognitive load, both in terms of invested mental effort and in terms of perceived task difficulty. We hypothesized that in line with the findings by Van Gog et al. (2012), single delayed ratings of mental effort after a sequence of problems would result in higher scores than the average of immediate multiple ratings provided immediately after each problem. What would happen with perceived difficulty ratings is an open question. As mentioned above, the concepts of invested mental effort and perceived task difficulty are to some extent related, in the sense that one would expect learners to invest more effort on tasks that they perceive to be more difficult (note that this may not apply at the extreme end of the scale, though; when tasks are perceived as too complex, learners might just give up and not invest any effort). Nevertheless, asking students to rate how much mental effort they invested in completing a task versus how difficult they perceived a task to be are two different questions that can lead to different interpretations. At least from a theoretical perspective, invested mental effort may involve more aspects than only the task (e.g., motivation), whereas perceived task difficulty seems to pertain mainly to the task itself (Van Gog and Paas 2008). Therefore, applying both measures may provide some more indications regarding the cause of the findings by Van Gog et al. (2012), and applying measures of interest or motivation could indicate whether the findings by Van Gog et al. regarding effort may have been caused by for instance motivational aspects, rather than experienced load. When the finding that a single rating at the end of a series of tasks is higher than the average of ratings per task is replicated for both effort and difficulty measures, but not for other kinds of measures such as motivation or interest, then this would seem to be an indication that the perceptions of cognitive load imposed by "the learning task", when making a single delayed rating over a series of tasks, differs from the average of multiple tasks.

Experiment 1

Participants

One hundred and sixty-eight German undergraduate students of educational science participated in this experiment. Their mean age was 23.13 years ($SD = 4.43$), and 68.5 % were female.

Method

Materials

Six so-called "weekday-problems" were used (Sweller 1993; Van Gog et al. 2012). The intrinsic cognitive load increased over the six problems in that we aimed at varying task complexity by increasing the element interactivity of the problems. For instance, the first problem was: "Suppose today is Tuesday. What day of the week is the day after tomorrow?" (low element interactivity), whereas the last problem was: "Suppose last Tuesday was the 2nd day. What day of the week is it in 17 days, if the 8th day is in two days? "(high element interactivity). Overall, we assumed the first three problems to be lower in intrinsic cognitive load (i.e. low complexity) than the last three problems, which were intended to be rather high in intrinsic cognitive load (i.e. high complexity). All six problems can be found in the Appendix.

Cognitive load was assessed by means of two subjective rating scales. First, a slightly modified and translated version of the mental effort rating scale developed by Paas (1992) was used. Participants were asked how much mental effort they had invested in solving each problem, ranging from very low (1) to very high (7). Second, the perceived task difficulty 7-point rating scale (cf. Kalyuga et al. 1999; Marcus et al. 1996) was used. Participants were asked how difficult they perceived each problem to be, ranging from very easy (1) to very difficult (7).

In addition, we also assessed perceived interest and motivation by using two items from a German version of the QCM (A questionnaire to assess current motivation in learning situations; Rheinberg et al. 2001) immediately after each problem as well as once at a delay (after all problems). Participants were asked to answer the items "I like such puzzles and riddles" as well as "I would work on such problems in my freetime", again on 7-point rating scales.

It should be noted that the original mental effort rating scale developed by Paas (1992) is answered on a 9-point rating scale. However, since both the perceived difficulty scale (Kalyuga et al. 1999) and the QCM items (Rheinberg et al. 2001) are answered on 7-point rating scales, and because these have been used for effort measures as well (see the review by Van Gog and Paas 2008), we decided to apply this shorter scale to our effort item as well to make the answer format more coherent for participants and to make the results more comparable. The two cognitive load items, perceived interest and motivation were assessed directly after each problem (immediate ratings) as well once after the whole sequence of problems (delayed rating).

Procedure

A repeated-measures-within-subjects-design was used. Students were tested during a lecture on educational sciences. They were instructed to solve the sequence of six problems,

each of which was presented on one Power Point slide, and they were given one minute to answer each problem. The time was kept by the instructor and students were not allowed to continue before time was up and the instructor moved on to the next slide. Students were asked to solve each problem without taking notes (but to write down their solution for each problem on the questionnaire). After solving each problem, they were instructed to rate their invested mental effort and perceived task difficulty immediately (immediate ratings). In addition, after having completed all six problems, they were asked to rate their overall invested mental effort and perceived task difficulty as well as motivation and interest in one delayed rating. Finally, students completed a short demographic questionnaire.

Results and discussion

General findings: do immediate and delayed measures really differ, and if so, what are the delayed ratings based on?

Means and standard deviations for problem performance, perceived task difficulty, invested mental effort, interest and motivation can be found in Table 1.

First, we calculated performance scores by assigning 1 point for each correctly solved problem and calculating an overall score for the sum of the six problems (which, accordingly, could vary between 0 and 6). As a check on the increase in problem complexity, we then conducted repeated measures analyses of variance, which indeed indicated a significant linear decrease in performance during problem solving, $F(5, 163) = 183.36$, $p < 0.001$, partial eta$^2 = 0.85$ (with the exception of the fourth problem which seemed to be a bit out of line in that performance on this problem was lower than on any other problem, all p's < 0.001), a significant linear increase in invested mental effort, $F(5, 163) = 355.33$, $p < 0.001$, partial eta$^2 = 0.92$, and a significant linear increase in perceived task difficulty, $F(5, 163) = 211.16$, $p < 0.001$, partial eta$^2 = 0.87$. In line with this, two paired samples t-tests revealed that students' summarized performance on problems a, b, and c ($M = 2.80$, $SD = 0.45$) was significantly higher than their performance on problems d, e, and f ($M = 1.02$, $SD = 0.95$), $t(167) = 22.80$, $p < 0.001$, $d = 2.39$. The same applied to the average of mental effort ratings for problems a, b, and c ($M = 1.95$, $SD = 0.61$), which was significantly lower than on problems d, e, and f ($M = 4.90$, $SD = 1.21$), $t(167) = -37.72$, $p < 0.001$, $d = 3.08$. Finally, average ratings of perceived task difficulty on problems a, b, and c ($M = 1.62$, $SD = 0.53$) were significantly lower than on problems d, e, and f ($M = 4.31$, $SD = 1.30$), $t(167) = -30.34$, $p < 0.001$, $d = 2.68$.

In short, the performance data show that the six problems are becoming more complex and are indeed perceived as becoming more and more difficult, requiring more and more effort to be invested. Moreover, in line with our intended manipulation of complexity, the first three problems appear to be lower in complexity than the last three problems as can be seen in the individual ratings for each of the six problems, in which there seems to be a gap between the ratings for the first three and for the last three items.

In a next step and according to the starting point for conducting this experiment, we were interested in whether the delayed scores given after the whole problem solving sequence differ from the average of the ratings after each task. We thus calculated mean scores for the respective six immediate ratings regarding invested mental effort, perceived task difficulty, interest and motivation and compared them to the respective delayed ratings. While we found that the average of ratings and the delayed ratings for mental effort and perceived difficulty all correlated with each other significantly (all r's > 0.56, all

Table 1 Experiment 1 ($N = 168$): means and standard deviations for problem solving performance, invested mental effort, perceived task difficulty, interest and motivation for each of the six problems as well as for the delayed measures

Problem	Performance		Invested mental effort		Perceived task difficulty		Interest		Motivation	
	M	*SD*	M	*SD*	M	*SD*	M	*SD*	M	*SD*
Problem a (1st problem)	1.000	0.00	1.107	0.347	1.077	0.268	3.649	1.775	2.548	1.747
Problem b (2nd problem)	0.947	0.226	1.762	0.897	1.429	0.688	3.631	1.773	2.566	1.658
Problem c (3rd problem)	0.857	0.351	2.994	1.035	2.345	1.032	3.589	1.779	2.691	1.706
Problem d (4th problem)	0.185	0.389	4.917	1.369	4.339	1.615	3.387	1.828	2.750	1.781
Problem e (5th problem)	0.417	0.494	4.637	1.534	4.066	1.560	3.310	1.834	2.691	1.801
Problem f (6th problem)	0.423	0.495	5.149	1.535	4.524	1.645	3.280	1.964	2.726	1.875
Delayed measure	0.498 (av.)	0.163	4.321	1.240	3.744	1.163	3.417	1.868	2.739	1.816

Table 2 Experiment 1 ($N = 168$): results of stepwise regression analyses for the prediction of the delayed mental effort rating from the immediate ratings

Model		Unstandardized coefficients		Standardized coefficients	T	*Sig.*
		B	*SE*	β		
1	(Constant)	1.859	0.231		8.051	<0.001
	Problem e (5th problem)	0.531	0.047	0.657	11.228	<0.001
2	(Constant)	1.152	0.256		4.492	<0.001
	Problem e (5th problem)	0.389	0.052	0.482	7.460	<0.001
	Problem f (6th problem)	0.265	0.052	0.328	5.077	<0.001
3	(Constant)	0.887	0.265		3.348	0.001
	Problem e (5th problem)	0.321	0.056	0.397	5.774	<0.001
	Problem f (6th problem)	0.245	0.051	0.303	4.771	<0.001
	Problem c (3rd problem)	0.229	0.075	0.191	3.061	0.003
4	(Constant)	0.664	0.283		2.348	0.020
	Problem e (5th problem)	0.285	0.058	0.353	4.956	<0.001
	Problem f (6th problem)	0.231	0.051	0.285	4.501	<0.001
	Problem c (3rd problem)	0.168	0.080	0.140	2.111	0.036
	Problem d (4th problem)	0.131	0.062	0.145	2.104	0.037

p's < 0.001), two paired samples t-tests revealed significant differences between the average of ratings after each task versus single delayed ratings of both invested mental effort ($M = 3.43$, $SD = 0.81$ vs. $M = 4.32$, $SD = 1.24$, $t(167) = 13.54$, $p < 0.001$, $d = 0.85$), and perceived task difficulty ($M = 2.96$, $SD = 0.81$ vs. $M = 3.74$, $SD = 1.16$, $t(167) = 12.28$, $p < 0.001$, $d = 0.78$). In contrast, regarding interest ($M = 3.47$, $SD = 1.66$ vs. $M = 3.42$, $SD = 1.87$, $t(167) < 1$) and motivation ($M = 2.66$, $SD = 1.65$ vs. $M = 2.74$, $SD = 1.82$, $t(167) = 1.64$, $p = 0.103$, $d = 0.04$), no significant differences were found between the average of ratings after each task and the single delayed rating.

It thus seems that there is something different about giving subjective ratings of cognitive load than of affective variables such as interest and motivation. To investigate potential reasons for the differences between delayed and immediate effort and difficulty ratings, we conducted two step-wise linear regression analyses that aimed at finding out which of the respective immediate ratings prove to be the strongest predictors for the delayed ratings. The results of these analyses are depicted in Tables 2 and 3.

As can be seen in the tables, there are four possible models including different combinations of predictors both for invested mental effort and perceived difficulty. The strongest single predictor for the delayed mental effort rating seems to be the fifth rating, $t(167) = 11.23$, $p < 0.001$, $r^2 = 0.432$. However, when taking a model into account that includes the fifth and sixth rating as predictors, the model is significant as well, with a change in r^2 of 0.077 (leading to $r^2 = 0.508$ for the model). The same applies for the third and fourth model that take into account ratings 3, 5 and 6 ($r^2 = 0.526$) and 3, 4, 5 and 6, respectively ($r^2 = 0.536$). Within these four models, the combination of immediate mental effort ratings always predicts the delayed mental effort rating significantly. Overall, it appears that for mental effort the latter and at the same time more complex problems/ratings contribute more to the delayed rating than the first and less complex ones.

The same analyses were conducted for perceived difficulty. The strongest single predictor for the delayed difficulty rating appeared to be the sixth rating, $t(167) = 9.49$,

Table 3 Experiment 1 ($N = 168$): results of stepwise regression analyses for the prediction of the delayed perceived difficulty from the immediate ratings

Model		Unstandardized coefficients		Standardized coefficients	*T*	*Sig.*
		B	*SE*	β		
1	(Constant)	1.847	0.213		8.685	<0.001
	Problem f (6th problem)	0.419	0.044	0.593	9.490	<0.001
2	(Constant)	1.201	0.216		5.571	<0.001
	Problem f (6th problem)	0.292	0.044	0.414	6.600	<0.001
	Problem e (5th problem)	0.300	0.047	0.402	6.420	<0.001
3	(Constant)	1.000	0.222		4.506	<0.001
	Problem f (6th problem)	0.262	0.045	0.370	5.870	<0.001
	Problem e (5th problem)	0.231	0.052	0.310	4.482	<0.001
	Problem d (4th problem)	0.143	0.049	0.198	2.903	0.004
4	(Constant)	0.813	0.233		3.487	0.001
	Problem f (6th problem)	0.270	0.044	0.382	6.119	<0.001
	Problem e (5th problem)	0.212	0.051	0.285	4.125	<0.001
	Problem d (4th problem)	0.120	0.050	0.166	2.420	0.017
	Problem b (2nd problem)	0.228	0.097	0.135	2.343	0.020

$p < 0.001$, $r^2 = 0.352$. However, and similar to the mental effort results, when taking a model into account that includes the fifth and sixth rating as predictors, the model is significant as well, with a change in r^2 of 0.130 (leading to $r^2 = 0.481$ for the model). The same applies for the third and fourth model that take into account ratings 4, 5 and 6 ($r^2 = 0.507$) and 2, 4, 5 and 6, respectively ($r^2 = 0.523$). Within these four models, the combination of immediate mental effort ratings always predicts the delayed mental effort rating significantly. Overall (with a slight exception with the 2nd rating in model 4), it again appears that for perceived difficulty, the latter and at the same time more complex problems/ratings contribute more to the delayed rating than the first and less complex ones.

How do the measures relate to performance?

Table 4 shows the results of regression analyses that were conducted to find out whether the measures for mental effort, perceived difficulty, interest and motivation predict performance at all. As can be seen, for mental effort, interest and motivation, both the averages of ratings after each task as well as the single delayed ratings were able to predict performance significantly. However, this did not apply to the difficulty ratings, with the exception of the rating on the sixth problem, which predicted performance on this problem, $t(167) = -2.13$, $p = 0.035$, $r^2 = 0.027$, but was not significant in predicting overall performance, $t(167) = -1.44$, $p = 0.151$, $r^2 = 0.012$.

Results summary

To sum up, our first findings revealed that with the exception of perceived difficulty, cognitive and affective variables such as mental effort, interest and motivation are suitable to predict problem performance and are thus worth being investigated in more detail. This is especially interesting, since for mental effort, both the average of ratings after each task

Table 4 Predicting performance from the average of immediate ratings as well as from the delayed ratings of mental effort, perceived task difficulty, interest and motivation: results of linear regression analyses ($N = 168$, Experiment 1)

	Unstandardized coefficients		Standardized coefficients		
	b	*SE*	*β*	*T*	*p*
Mental Effort: av. of immediate ratings	−0.235	0.104	−0.174	−2.275	**0.024**
Mental effort: delayed rating	−0.135	0.068	−0.152	−1.986	**0.049**
Perceived difficulty: av. of immediate ratings	−0.150	0.105	−0.110	−1.422	0.157
Perceived difficulty: delayed rating	−0.073	0.073	−0.077	−0.993	0.322
Interest: av. of immediate ratings	0.139	0.050	0.210	2.768	**0.006**
Interest: delayed rating	0.160	0.044	0.271	3.631	**<0.001**
Motivation: av. of immediate ratings	0.145	0.050	0.218	2.874	**0.005**
Motivation: delayed rating	0.150	0.046	0.247	3.286	**0.001**

Bold values stand for significant results

as well as the delayed measure, predicted performance significantly, but at the same time, these two measures differed significantly from each other. The question thus remains what these two measures really reflect. Our first findings might suggest that delayed ratings for both cognitive load measures, which are significantly higher than the average of immediate ratings, are based on the most complex problems in a series of problems. However, given that in this experiment, the order of the problems was fixed, it is possible that this finding at least partly resulted from the specific order in which the problems were presented (i.e., a recency effect; e.g., Logie 1995). To shed more light on this possible confounding of complexity and order of presentation, we conducted a second experiment, in which the order of the problems was varied.

Experiment 2

Participants

One hundred and seven German undergraduate students of educational science participated in this experiment (none of them had participated in Experiment 1). Their mean age was 24.8 years ($SD = 4.2$), and 69.2 % were female.

Method

Materials

The same weekday problems and cognitive load rating scales as in Experiment 1 were used (Appendix). Again the intrinsic cognitive load was systematically varied through the different complexities of the six problems. However, in this experiment we used two sequences of problems. In the first sequence the intrinsic cognitive load systematically increased over the course of the six problems, so this sequence was similar to Experiment 1 with one exception: Based on the results of Experiment 1, we changed the position of problem d, which was moved from the fourth to the fifth position, and problem e was

moved from the fifth to the fourth position. Thus, the first sequence of problems was a, b, c, e, d, f. The first three problems were again assumed to be lower in intrinsic cognitive load (i.e. low complexity) than the last three problems, which were assumed to be higher in intrinsic cognitive load (i.e. high complexity). In the second sequence the intrinsic cognitive load systematically decreased over the course of the six problems (i.e. the order was reversed, resulting in the sequence f, d, e, c, b, a). That is, in the second sequence, the first three problems were higher in intrinsic cognitive load (i.e. high complexity) than the last three problems, which were lower in intrinsic cognitive load (i.e. low complexity).

Procedure

Similar to the first experiment, a repeated-measures-between-subjects-design was used. Students were again tested during lectures on educational sciences. They were randomly assigned to one of the two sequences of problems (Group 1: increasing complexity, $n = 56$; Group 2: decreasing complexity, $n = 51$). Instructions and time-keeping procedure were the same as in Experiment 1.

Results and discussion

General findings

Means and standard deviations for problem performance, perceived task difficulty, invested mental effort, interest and motivation for the two groups receiving different sequences of the six problems can be seen in Tables 5 and 6. Similar to Experiment 1, performance scores were again calculated by assigning 1 point for each correctly solved problem and calculating an overall score for the sum of the six problems (which, accordingly, could vary between 0 and 6).

Again, to check whether problem complexity and thus intrinsic load can be assumed to increase respectively decrease for the two groups, we conducted repeated measures analyses of variance. For performance, it was indeed shown that a significant linear decrease occurred for Group 1, $F(5, 51) = 101.03$, $p < 0.001$, partial $\text{eta}^2 = 0.908$, while a significant linear increase occurred for Group 2, $F(5, 46) = 71.19$, $p < 0.001$, partial $\text{eta}^2 = 0.886$. Analogously, mental effort increased linearly for Group 1, $F(5, 51) = 281.37$, $p < 0.001$, partial $\text{eta}^2 = 0.965$, and decreased linearly for Group 2, $F(5, 46) = 154.96$, $p < 0.001$, partial $\text{eta}^2 = 0.944$. Finally, the difficulty ratings also showed a significant linear increase for Group 1, $F(5, 51) = 127.22$, $p < 0.001$, partial $\text{eta}^2 = 0.926$ and a significant linear decrease for Group 2, $F(5, 46) = 59.22$, $p < 0.001$, partial $\text{eta}^2 = 0.866$.

In addition, similar to Experiment 1, we calculated paired samples t-tests with which we compared the summed performance and averaged cognitive load scores for the first three versus the last three problems, which we assumed to be rather easy and rather complex, respectively (depending on the sequence in which they were presented). These findings again support the findings on the repeated measures analyses. For Group 1, the summed performance on the first three problems ($M = 2.66$, $SD = 0.58$) was significantly higher than on the last three problems ($M = 0.70$, $SD = 0.81$), $t(55) = 18.72$, $p < 0.001$, $d = 2.82$. The average mental effort rating for the first three problems ($M = 1.90$, $SD = 0.54$) was significantly lower than for the last three problems ($M = 5.22$, $SD = 0.85$), $t(55) = -32.37$, $p < 0.001$, $d = 4.71$, and the same applied to the average difficulty rating for the first three problems ($M = 1.65$, $SD = 0.41$) versus the last three

Table 5 Experiment 2, Group 1 ($N = 56$): means and standard deviations for problem solving performance, invested mental effort, perceived task difficulty, interest and motivation for each of the six problems as well as for the delayed measures for participants who received problems *increasing* in complexity

Problem	Performance		Invested mental effort		Perceived task difficulty		Interest		Motivation	
	M	*SD*	M	*SD*	M	*SD*	M	*SD*	M	*SD*
Problem a (1st problem)	0.982	0.134	1.161	0.417	1.054	0.227	3.786	1.875	2.839	1.876
Problem b (2nd problem)	0.786	0.414	1.759	0.580	1.393	0.528	3.782	1.873	2.709	1.760
Problem c (3rd problem)	0.893	0.312	2.786	0.889	2.518	0.894	3.589	1.817	2.768	1.768
Problem e (4th problem)	0.357	0.483	4.553	1.060	4.018	1.328	3.393	1.744	2.679	1.562
Problem d (5th problem)	0.107	0.312	5.464	1.250	4.964	1.439	3.429	1.818	2.696	1.683
Problem f (6th problem)	0.232	0.426	5.643	1.151	5.232	1.465	3.286	1.885	2.750	1.771
Delayed measure	0.560 (av.)	0.194	4.339	1.116	3.964	1.175	3.375	1.874	2.786	1.681

Table 6 Experiment 2, Group 2 ($N = 51$): means and standard deviations for problem solving performance, invested mental effort, perceived task difficulty, interest and motivation for each of the six problems as well as for the delayed measures for participants who received problems *decreasing* in complexity

Problem	Performance		Invested mental effort		Perceived task difficulty		Interest		Motivation	
	M	*SD*	M	*SD*	M	*SD*	M	*SD*	M	*SD*
Problem f (1st problem)	0.216	0.415	5.431	1.118	4.333	1.532	3.647	1.842	2.961	1.897
Problem d (2nd problem)	0.177	0.385	5.333	1.337	4.628	1.637	3.706	1.973	2.940	1.984
Problem e (3rd problem)	0.275	0.451	4.588	1.186	3.882	1.351	3.451	1.836	2.922	1.968
Problem c (4th problem)	0.745	0.440	3.510	1.173	2.863	1.114	3.373	1.865	2.922	1.937
Problem b (5th problem)	0.927	0.272	2.275	0.874	1.843	0.834	3.333	1.862	2.784	1.836
Problem a (6th problem)	1.000	0.000	1.157	0.418	1.137	0.401	2.961	1.788	2.784	1.847
Delayed measure	0.556 (av.)	0.191	4.255	0.891	3.745	0.935	3.431	1.814	2.902	1.911

problems ($M = 4.74$, $SD = 1.11$), $t(55) = -22.68$, $p < 0.001$, $d = 3.72$. For Group 2, these comparisons revealed that on the first three problems ($M = 0.67$, $SD = 0.89$) performance was significantly lower than on the last three problems ($M = 2.67$, $SD = 0.59$), $t(50) = -14.57$, $p < 0.001$, $d = 2.68$. The average mental effort scores for the first three problems ($M = 5.12$, $SD = 0.90$) were significantly higher than for the last three problems ($M = 2.31$, $SD = 0.68$), $t(50) = 20.06$, $p < .001$, $d = 3.53$. Finally, the perceived difficulty ratings for the first three problems ($M = 4.28$, $SD = 1.25$) were significantly higher than for the last three problems ($M = 1.95$, $SD = 0.67$), $t(50) = 16.70$, $p < 0.001$, $d = 2.35$.

To sum up, and again in line with our intended manipulation of the problem complexity, the six problems appear to become more and more difficult, requiring more effort to be invested for the first group; while for the second group, they become easier, requiring less effort.

Also similar to Experiment 1, our next step was to investigate whether the delayed scores given after the whole problem solving sequence differ from the average of the immediate ratings. As we were interested in whether this comparison differs for the two groups who receive the six problems in the original versus reversed order, we first calculated eight paired samples t-tests this time, two for each group. Similar to Experiment 1, the delayed ratings for interest and motivation did not differ from the average of ratings after each task in either Group 1 or 2 (all p's > 0.10). For mental effort, however, the delayed rating was significantly higher than the average rating for Group 1 ($M = 4.34$, $SD = 1.17$ vs. $M = 3.56$, $SD = 0.60$), $t(55) = 5.69$, $p < .001$, $d = 0.88$ as well as Group 2 ($M = 4.25$, $SD = 0.89$ vs. $M = 3.72$, $SD = 0.63$), $t(50) = 5.85$, $p < 0.001$, $d = 0.71$. Regarding perceived difficulty, the delayed rating for Group 1 ($M = 3.96$, $SD = 1.17$) was significantly higher than the average of ratings after each task ($M = 3.20$, $SD = 0.66$), $t(55) = 6.33$, $p < 0.001$, $d = 0.81$, and the same applied for Group 2 (delayed: $M = 3.75$, $SD = 0.93$; average: $M = 3.11$, $SD = 0.87$), $t(50) = 5.57$, $p < 0.001$, $d = 0.71$.

These findings for the two groups also remained constant when taking the whole sample of participants into account. While neither the delayed versus immediate scores for interest ($t(106) = -1.29$, $p = 0.200$) nor for motivation ($t < 1$) differed, there were overall differences for the cognitive load ratings. The delayed rating of invested mental effort ($M = 4.29$, $SD = 1.01$) was significantly higher than the average of the six immediate ratings ($M = 3.63$, $SD = 0.61$), $t(106) = -7.87$, $p < .001$, $d = 0.79$. Similarly, the delayed rating of perceived task difficulty ($M = 3.86$, $SD = 1.07$) was significantly higher than the average of the six immediate ratings ($M = 3.16$, $SD = 0.76$), $t(106) = -8.44$, $p < 0.001$, $d = 0.75$. In addition, and also in line with Experiment 1, all immediate and delayed measures for mental effort and perceived difficulty correlated significantly with each other (all r's > 0.31, all p's < 0.01).

In sum, results of Experiment 1 were replicated: The delayed cognitive load ratings, that is, mental effort and perceived difficulty, were higher than the averages of ratings after each task, whereas there was no such difference for affective variables, that is, interest and motivation.

The results of Experiment 1 suggested that the single delayed ratings for mental effort and perceived difficulty were predicted by ratings on the more complex problems, that is, the ones that were presented later in the sequence. To shed more light on the question whether these findings can really be traced back to task complexity or rather to some kind of recency effect, we conducted step-wise linear regression analyses for the two groups separately to find out which of the immediate ratings best predict the respective delayed rating. If the differences can really be based on the participants' orientation towards the

Table 7 Experiment 2 ($N = 107$): Results of stepwise regression analyses for the prediction of the delayed mental effort rating from the immediate ratings, separately shown for the group receiving problems increasing in complexity ($N = 56$) and decreasing in complexity ($N = 51$)

Model		Unstandardized coefficients		Standardized coefficients	T	*Sig.*
		B	*SE*	*β*		
Increasing complexity						
1	(Constant)	2.463	0.616		3.997	<0.001
	Problem e (4th problem)	0.412	0.132	0.391	3.125	0.003
Decreasing complexity						
1	(Constant)	2.341	0.423		5.539	<0.001
	Problem d (2nd problem)	0.417	0.089	0.556	4.676	<0.001
2	(Constant)	1.911	0.431		4.432	<0.001
	Problem d (2nd problem)	0.325	0.091	0.433	3.561	0.001
	Problem c (4th problem)	0.243	0.092	0.320	2.630	0.011

Table 8 Experiment 2 ($N = 107$): results of stepwise regression analyses for the prediction of the delayed perceived difficulty rating from the immediate ratings, separately shown for the group receiving problems increasing in complexity ($N = 56$) and decreasing in complexity ($N = 51$)

Model		Unstandardized coefficients		Standardized coefficients		
		B	*SE*	*β*	T	*Sig.*
Increasing complexity						
1	(Constant)	1.771	0.505		3.504	0.001
	Problem f (6th problem)	0.419	0.093	0.523	4.505	0.001
Decreasing complexity						
1	(Constant)	2.535	0.318		7.980	<0.001
	Problem d (2nd problem)	0.423	0.104	0.504	4.080	<0.001
2	(Constant)	2.007	0.363		5.530	<0.001
	Problem d (2nd problem)	0.293	0.110	0.349	2.663	0.011
	Problem e (3rd problem)	0.194	0.075	0.340	2.597	0.012

more complex problems, the findings for Group 1 should be similar to the findings from the first experiment, whereas the findings for Group 2 should be just the other way round. The results for the step-wise regressions for mental effort and difficulty for both groups can be found in Tables 7 and 8.

As can be seen in the tables, the results for these analyses are not as clear as they were in Experiment 1. For Group 1 who received problems with increasing complexity (that is, a procedure similar to Experiment 1), analyses revealed one model, which states that the strongest predictor for the delayed mental effort rating is the rating on the fourth problem, $t(55) = 3.13$, $p = 0.003$, $r^2 = 0.153$. With regard to the perceived difficulty rating, analyses revealed one model as well for this group, which states that the best predictor for the delayed difficulty rating is the rating on the sixth problem, $t(55) = 4.51$, $p < .001$, $r^2 = 0.273$. Therefore, in short, the delayed ratings for mental effort as well as for difficulty were best predicted by two of the problems that came later in the sequence and that we assume to belong to the more complex ones. So this finding is in line with the findings from the first experiment.

For Group 2, who received the problems in reversed order and thus with decreasing complexity, starting with the most complex problem, the analyses revealed two possible models for mental effort as well as for perceived difficulty. The strongest single predictor for the delayed mental effort rating seemed to be the second rating, $t(50) = 4.68$, $p < 0.001$, $r^2 = 0.309$. However, taking a model into account that includes the second and the fourth problem, would reveal a significant prediction as well with a change in r^2 of 0.087 (leading to $r^2 = 0.396$ for this model). With regard to perceived difficulty, one possible model again shows the second problem as the strongest single predictor for the delayed rating, $t(50) = 4.08$, $p < 0.001$, $r^2 = 0.254$. However, a second model including the second and the third problem would also predict the delayed difficulty rating with a change in r^2 of 0.092 (leading to $r^2 = 0.346$ for that model). To sum up the findings for the second group, who received the problems in reversed order (from complex to easy), it was found that three out of the four possible step-wise regression models included problems that participants received in the first half of the sequence, and thus problems that we assume to be more complex ones. An exception is the second model for the prediction of mental effort, where the fourth problem presented contributes as a predictor. This problem was actually assumed to be an easier one (and performance, effort, and difficulty data pointed in this direction as well). To shed more light on this issue, we averaged the ratings for mental effort and perceived difficulty regarding the first and the last three problems and calculated regression analyses again. For Group 1, these analyses revealed that the average of the first three problems (which we assume to be the easier ones) neither predicted the delayed mental effort rating, $t(55) = 1.17, p > 0.20$ nor the delayed difficulty rating, $t < 1$. In contrast, the delayed ratings for mental effort and perceived difficulty were significantly predicted by the average of the last three problems ($t(55) = 3.25$, $p = 0.002$, $r^2 = 0.164$ for mental effort and $t(55) = 5.97$, $p < 0.001$, $r^2 = 0.398$ for perceived difficulty), which we assume to be the more complex ones. The average of these three complex problems also predicts the delayed ratings for mental effort, $t(50) = 5.02$, $p < 0.001$, $r^2 = 0.340$ and perceived difficulty, $t(50) = 5.03, p < 0.001, r^2 = 0.348$ for Group 2. For this group, the average of the three easier problems did not predict the delayed difficulty score, $t(50) = 1.26, p > 0.20$; however, it did predict the delayed mental effort score, $t(50) = 2.16, p = 0.036, r^2 = 0.219$.

Taken together, the analyses for Experiment 2 revealed that for Group 1, who received the problems in the same order as the participants in the first experiment, the results are comparable to those in the first experiment. That is, the delayed cognitive load scores were best predicted by the problems that participants received later in the sequence and that we assume to be the more complex ones. For Group 2, who received the problems in reversed order from complex first to easy last, a slightly different picture emerges. While overall, the delayed cognitive load scores for this group appear to be predicted best by the problems they received earlier in the sequence (that is, also the more complex ones), at least the delayed mental effort rating can also be predicted by the average of the easy problems (that is, the three problems they receive last in the sequence).

How do the measures relate to performance?

Table 9 shows the results of regression analyses that were conducted to find out whether the measures for mental effort, perceived difficulty, interest and motivation predict performance at all. As can be seen, the predictive values of the cognitive load scores are quite in contrast to those in Experiment 1, because this time, neither the average of ratings after each task, nor the single delayed ratings after all tasks predicted performance. (Although the analyses revealed that $t(106) = -1.69$ and $p = 0.094$ for the average of immediate difficulty ratings and that $t(106) = -1.91$ and $p = 0.059$ for the delayed rating, these

Table 9 Predicting performance from the average of immediate ratings as well as from the delayed ratings of mental effort, perceived task difficulty, interest and motivation: Results of linear regression analyses ($N = 107$, Experiment 2)

	Unstandardized coefficients		Standardized coefficients		
	b	*SE*	*β*	*T*	*p*
Mental effort: av. of immediate ratings	0.126	0.182	0.067	0.691	0.491
Mental effort: delayed rating	−0.001	0.111	−0.001	−0.005	0.996
Perceived difficulty: av. of immediate ratings	−0.244	0.145	−0.163	−1.688	*0.094*
Perceived difficulty: delayed rating	−0.197	0.103	−0.183	−1.906	*0.059*
Interest: av. of immediate ratings	0.156	0.065	0.227	2.389	**0.019**
Interest: delayed rating	0.115	0.060	0.184	1.915	*0.058*
Motivation: av. of immediate ratings	0.145	0.050	0.218	2.874	**0.005**
Motivation: delayed rating	0.110	0.066	0.160	1.663	*0.099*

Bold values stand for significant results, italics values in stand for marginal results (i.e. *p* values between 0.05 and 0.1)

effects are not statistically significant and can therefore not be interpreted unambiguously.) The immediate and delayed scores for interest and motivation also did predict performance; however, the effects especially for the delayed scores were not as strong as they had been in the first experiment. This pattern of results also remains when having a look at the two separate groups, who received the problems in different orders, but the effects were smaller for all analyses, presumably due to the smaller samples.

Results summary

Our findings in the second experiment again let us assume that cognitive and affective variables seem to be suitable to predict problem performance, although the effects did not show up for the mental effort ratings this time and failed to reach statistical significance for difficulty slightly. Similar to Experiment 1, all cognitive load ratings correlated significantly with each other, but at the same time, the respective scores for the delayed ratings versus the average of ratings after each task, differed significantly from each other. Also in line with the results of Experiment 1, the results in Experiment 2 largely seem to confirm the assumption that the delayed ratings are based more on the more complex problems in the series and less on a recency effect. The overall conclusions that can be drawn from the findings of the two studies will be discussed in the next section (Table 9).

General discussion

The purpose of the present study was to investigate whether measures of the cognitive variables mental effort and difficulty that are collected with rating scales are suitable to predict performance and whether there is a difference between the average of multiple ratings that are given immediately during learning or problem solving versus a rating that is given just once after a learning or problem-solving phase. Furthermore, we added measures of affective variables (such as interest and motivation) to see whether these are also estimated differently during versus after learning and problem solving; if not, this suggests there is something different about cognitive load measures compared to affective measures.

We replicated the finding by Van Gog et al. (2012), that a single delayed mental effort rating at the end of a series of tasks is higher than the average of ratings on those tasks, and that this effect is independent of the sequence of problems. Furthermore, we extended their findings. First, while we used a similar type of problem, we used tasks at six complexity levels (instead of two), so we replicated the finding even though our series of tasks was different. Second, and more interestingly, we showed that the same finding applied to another index associated with cognitive load, namely difficulty ratings, but not to affective ratings of motivation or interest.

The reported result that a single delayed mental effort rating at the end of a series of tasks is higher than the average of ratings on those tasks is mainly relevant when such measures relate to performance. In our two studies, this was the case, although the regression analyses revealed mixed results. Whereas in Experiment 1, mainly the two mental effort ratings (immediate and delayed) predicted problem performance, while the respective results for the regression analyses regarding perceived difficulty did not reach statistical significance, it was the other way round in Experiment 2, where mainly effects for immediate and delayed difficulty scores could be found, while the regression results for the two mental effort scores missed statistical significance. That is, we found that mental effort predicted performance in Experiment 1 and that perceived difficulty predicted performance in Experiment 2, but unfortunately, in neither of the two experiments, both dimensions of cognitive load predicted performance at the same time. These mixed findings are also reflected in the results of the analyses we conducted when summarizing the immediate mental effort and perceived difficulty ratings for the three easy and the three complex problems, respectively and relating them to performance.

With regard to the question, why there are differences between the immediate and delayed ratings in both experiments and all other studies conducted so far, our results suggest the assumption that the perception of "the task" as a series of problems rather than individual problems, affects cognitive load ratings but not affective ratings. That is, when it comes to affective variables, the interest in certain kinds of problems and the motivation to solve such problems is probably based more on the *characteristics* of the problems themselves and is less likely to be affected by the *number* of problems or the total length of the task (i.e., the series). That is, students might have liked or not liked the weekday problems in general, and although their interest in the problem also depended on problem difficulty (as can be seen in the results of the regression analyses), the gap between the ratings of problems at different complexity levels was not too large and the single delayed rating did not differ from the average of ratings on each task.

With regard to cognitive load ratings, while there are several indicators for the assumption that delayed cognitive load ratings are higher than the average of immediate ratings because they are based more on the students remembering the more complex problems they have received rather than on having all problems in mind when estimating their cognitive load after the whole learning or problem solving phase, some of our findings also make other interpretations possible. One reason for the mixed results might be that the students do not only take the most difficult problems into account, but also the number of immediate ratings they had to give and/or the number of problems they had to solve when giving their delayed rating. However, Van Gog et al. (2012) investigated this by a between-conditions variation of immediate and delayed scores, and showed that this could not explain the difference. More specifically, they compared two groups, which both received six weekday problems; whereby the order of the problems was varied within each group (participants either received three simple and then three complex problems or the other way round, or simple and complex problems alternatingly in succession). One group rated

their mental effort immediately after each task, but not delayed after the whole problem solving process, whereas the other group only did this one delayed rating. Results again were similar to our findings. Thus, as mentioned above, an explanation might rather relate to students' perception of "the learning task" as a whole; that is, when being asked to give delayed estimations of their cognitive load, participants take the total number of problems or the total time into account rather than simply remembering each (or the most difficult) problem and its characteristics. Future research should attempt to shed light on this issue for instance by means of mixed designs that work with different amounts of problems presented in varying sequences. Additionally, future research should investigate whether the effect also appears with other tasks than problem solving (e.g., text comprehension).

Nevertheless, based on the results of our studies, it does seem safe to suggest that researchers should be conscious of the fact that the time points at which they apply subjective CL rating scales will affect the data they obtain. One might argue that the difference shown between the average of immediate scores and a delayed score of cognitive load for subjective ratings is trivial. But given the amount of empirical studies dealing with memory, learning, thinking, or problem solving that use subjective CL measures to interpret effects of instructional design on learning, it is not. Thus, future research should aim to clarify what exactly causes these delayed ratings to be higher than the average of immediate ratings, and which application of cognitive load rating scales provides the most reliable (and valid?) indicator of the actual load experienced: immediate ratings given after every task in a learning or test phase; a delayed rating after the entire learning or test phase; or an average of both immediate and delayed ratings. This could potentially be achieved by comparing objective measures of CL such as dual-task measures (e.g., Brünken et al. 2004; Renkl et al. 2003) or neuro-physiological measures (e.g., Antonenko et al. 2010; Paas and van Merriënboer 1994; Van Gerven et al. 2004) with subjective ratings obtained at different points in time.

Appendix

Weekday problems (according to order of presentation in Experiment 1)

(a) Suppose today is Tuesday. What day of the week is the day after tomorrow?
(b) Suppose today is Thursday. What day of the week is in three days?
(c) Suppose yesterday was Wednesday. What day of the week was four days before the day before yesterday?
(d) What day was yesterday if the day after the day after tomorrow is three days before Sunday?
(e) Suppose five days after the day before yesterday is Friday. What day of the week is tomorrow?
(f) Suppose last Tuesday was the 2nd day. What day of the week is it in 17 days, if the 8th day is in two days?

References

Antonenko, P., & Niederhauser, D. (2010). The influence of leads on cognitive load and learning in a hypertext-assisted learning environment. *Computers in Human Behavior, 26*, 140–150.

Antonenko, P., Paas, F., Grabner, R., & Van Gog, T. (2010). Using electroencephalography (EEG) to measure cognitive load. *Educational Psychology Review, 22*, 425–438.

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic load within problems. *Learning and Instruction, 16*, 389–400.
Bratfisch, O., Borg, G., & Dornic, S. (1972). *Perceived item difficulty in three tests of intellectual performance capacity*. Stockholm: Institute of Applied Psychology, Report No. 29.
Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist, 38*, 53–61.
Brünken, R., Plass, J. L., & Leutner, D. (2004). Assessment of cognitive load in multimedia learning with dual-task methodology: Auditory load and modality effects. *Instructional Science, 32*, 115–132.
Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*, 293–332.
Chandler, P., & Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. *British Journal of Educational Psychology, 62*, 233–246.
Chandler, P., & Sweller, J. (1996). Cognitive load while learning to use a computer program. *Applied Cognitive Psychology, 10*, 1–20.
de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science, 38*, 105–134.
Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology, 13*, 351–371.
Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology, 93*, 579–588.
Kühl, T., Scheiter, K., Gerjets, P., & Edelmann, J. (2011). The influence of text modality on learning with static and dynamic visualizations. *Computers in Human Behavior, 27*, 29–35.
Kuusela, H., & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *American Journal of Psychology, 113*, 387–404.
Leutner, D., Leopold, C., & Sumfleth, E. (2009). Cognitive load and science text comprehension: Effects of drawing and mentally imagining text content. *Computers in Human Behavior, 25*, 284–289.
Logie, R. H. (1995). *Visuo-spatial working memory*. Hove, UK: Lawrence Erlbaum Associates.
Marcus, N., Cooper, M., & Sweller, J. (1996). Understanding instructions. *Journal of Educational Psychology, 88*, 49–63.
Opfermann, M. (2008). *There's more to it than instructional design: The role of individual learner characteristics for hypermedia learning*. Berlin, Germany: Logos.
Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*, 429–434.
Paas, F., Ayres, P., & Pachman, M. (2008). Assessment of cognitive load in multimedia learning: Theory, methods and applications. In D. Robinson & G. Schraw (Eds.), *Recent innovations in educational technology that facilitate student learning* (pp. 20–36). Charlotte, NC: Information Age Publishing.
Paas, F., Renkl, A., & Sweller, J. (2003a). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*, 1–4.
Paas, F., Tuovinen, J., Tabbers, H. K., & Van Gerven, P. W. M. (2003b). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*, 63–71.
Paas, F., & van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86*, 122–133.
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.
Park, B., Moreno, R., Seufert, T., & Brünken, R. (2011). Does cognitive load moderate the seductive details effect? A multimedia study. *Computers in Human Behavior, 27*, 5–10.
Plass, J. L., Moreno, R., & Brünken, R. (Eds.). (2010). *Cognitive load: Theory & application*. Cambridge: Cambridge University Press.
Renkl, A., Gruber, H., Weber, S., Lerche, T., & Schweizer, K. (2003). Cognitive Load beim Lernen aus Lösungsbeispielen [Cognitive load of learning from worked-out examples]. *Zeitschrift für Pädagogische Psychologie, 17*, 93–101.
Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen [QCM: A questionnaire to assess current motivation in learning situations]. *Diagnostica, 47*, 57–66.
Schwamborn, A., Thillmann, H., Opfermann, M., & Leutner, D. (2011). Cognitive load and instructionally supported learning with provided and learner-generated visualizations. *Computer in Human Behavior, 27*, 89–93.
Seufert, T., & Brünken, R. (2006). Cognitive load and the format of instructional aids for coherence formation. *Applied Cognitive Psychology, 20*, 321–331.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257–285.
Sweller, J. (1993). Some cognitive processes and their consequences for the organization and presentation of information. *Australian Journal of Psychology, 45*, 1–8.
Sweller, J. (2010). Element interactivity and intrinsic, extraneous and germane cognitive load. *Educational Psychology Review, 22*, 123–138.
Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York: Springer.
Sweller, J., van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296.
Tabbers, H. K., Martens, R. L., & van Merriënboer, J. J. G. (2004). Multimedia instructions and cognitive load theory: Effects of modality and cueing. *British Journal of Educational Psychology, 74*, 71–81.
Taylor, K. L., & Dionne, J.-P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 92*, 413–425.
Van Gerven, P. W. M., Paas, F., van Merriënboer, J. J. G., & Schmidt, H. G. (2004). Memory load and the cognitive pupillary response in aging. *Psychophysiology, 41*, 167–174.
Van Gog, T., & Jarodzka, H. (2013). Eye tracking as a tool to study and enhance cognitive and metacognitive processes in computer-based learning environments. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 143–156). New York: Springer.
Van Gog, T., Kirschner, F., Kester, L., & Paas, F. (2012). Timing and frequency of mental effort measurement: Evidence in favor of repeated measures. *Applied Cognitive Psychology, 26*, 833–839.
Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist, 43*, 16–26.
Van Gog, T., Paas, F., & van Merriënboer, J. J. G. (2006). Effects of process-oriented worked examples on troubleshooting transfer performance. *Learning and Instruction, 16*, 154–164.
Van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*, 147–177.