Timing and Frequency of Mental Effort Measurement: Evidence in Favour of Repeated Measures

Running head: MEASURING MENTAL EFFORT

Timing and Frequency of Mental Effort Measurement: Evidence in Favor of Repeated Measures

Tamara van Gog[a], Femke Kirschner[b] Liesbeth Kester[c], & Fred Paas[a,d]

[a] Institute of Psychology, Erasmus University Rotterdam, The Netherlands

[b] Department of Educational Sciences, Utrecht University, The Netherlands

[c] Centre for Learning Sciences and Technologies, Open University of The Netherlands

[d] Faculty of Education, University of Wollongong, Australia

Author Note:

Correspondence concerning this manuscript should be addressed to Tamara van Gog, Institute of Psychology, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. T: +31 10 4089041. E: vangog@fsw.eur.nl

Acknowledgement. During the realization of this work, Tamara van Gog was supported by a Veni grant from the Netherlands Organization for Scientific Research (NWO; # 451-08-003).

Abstract

Subjective mental effort rating scales are widely used in research on learning, instruction, and training. However, the timing and frequency of application of those rating scales differs between studies. Some apply a rating scale repeatedly after every task in a learning or test phase, others only once at the end of a phase. Four experiments are presented that investigated how timing and frequency of mental effort measurements affect the results obtained. The findings from Experiment 1 (between-subjects) and 2 (within-subjects), using different arrangements of simple and complex tasks, showed that a single rating after a series of tasks resulted in a higher mental effort score than the average of ratings provided immediately after every task. A similar result was obtained in Experiment 3 with series of complex tasks, but not with simple tasks. Experiment 4 showed that knowing beforehand that mental effort rating will be required after completing all tasks results in lower scores, but average retrospective ratings per task still differed from a single retrospective rating. Taken together, these experiments suggest that timing and frequency of effort ratings do affect the results obtained and that repeatedly measuring mental effort after each task in the series seems to be preferable.

Keywords: mental effort; subjective ratings; cognitive load; task complexity

Timing and Frequency of Mental Effort Measurement: Evidence in Favor of Repeated Measures

In research on learning, instruction, and training, subjective measures of mental or cognitive effort (Paas, 1992; Yeo & Neal, 2008) are regularly applied. Such measurements can reveal important additional information for researchers that is not necessarily reflected by more common performance measures such as accuracy, speed, or number/type of errors made. Particularly, the combination of performance and mental effort measures can provide information concerning the relative efficiency of training methods, in terms of the knowledge or skill acquisition process and the quality of knowledge or skills acquired (see e.g., Hoffman, & Schraw, 2010; Paas & Van Merriënboer, 1993; Van Gog & Paas, 2008), as well as motivational factors in training (see e.g., Kanfer & Ackerman, 1989; Paas, Tuovinen, Van Merriënboer, & Darabi, 2005; Yeo & Neal, 2004).

Different subjective mental effort rating scales are available. For example, the multidimensional NASA-Task Load Index (NASA-TLX; Hart & Staveland, 1988) incorporates measures of perceived performance, effort, frustration, and mental, physical and temporal task demands. In a modified form (to make it suitable for cognitive tasks) NASA-TLX has also been used in educational research (e.g., Gerjets, Scheiter, & Catrambone, 2006; Kester, Lehnen, Van Gerven, & P. A. Kirschner, 2006; Scheiter, Gerjets, Vollmann, & Catrambone, 2009). Yeo and Neal (2004, 2008) applied an 11-point single item rating scale asking participants to rate how hard they were trying, ranging from (0) not at all, to (10) extremely hard. Paas (1992) developed a uni-dimensional 9-point symmetrical rating scale which asks participants to rate how much mental effort they invested in a task, ranging from (1) very very low mental effort, to (9) very, very high mental effort. Mental effort has been defined by Paas and colleagues as “the cognitive capacity that is actually allocated to accommodate the demands imposed by the task; thus, it can be considered to reflect the actual cognitive load” (Paas, Tuovinen, Tabbers, & Van Gerven, 2003, p. 64). This measure or a slightly adapted version thereof has been applied in many educational psychology

studies (see Paas et al., 2003; Van Gog & Paas, 2008), particularly those inspired by cognitive load theory (Sweller, Ayres & Kalyuga, 2011; Sweller, Van Merriënboer, & Paas, 1998).

Although single-item scales have been criticized, Yeo and Neal (2008) discuss research that shows that for unambiguous constructs such as perceived mental effort or perceived task difficulty, single-item measures have psychometric properties equalling those of multi-item measures. Moreover, Paas's 9-point mental effort rating scale has been shown to be sensitive enough to detect small variations in task complexity (see Paas, Van Merriënboer, & Adam, 1994; Van Gerven, Paas, Van Merriënboer, & Schmidt, 2004).

An important question, however, is when to apply such subjective rating scales, as researchers tend to apply them either multiple times during or once after the learning or test phase. The first type of studies asks participants to provide a rating immediately after every task in a learning or test phase (e.g., single-item scale: Kirschner, Paas, & Kirschner, 2009; Kirschner, Paas, Kirschner, & Janssen, 2011; Nievelstein, Van Gog, Van Dijck, & Boshuizen, 2011; Paas, 1992; Paas, & Van Merriënboer, 1994; Spanjers, Wouters, Van Gog, & Van Merriënboer, 2011; Van Gog, Kester, & Paas, 2011; Van Gog, Paas, & Van Merriënboer, 2006, 2008; Van Merriënboer et al., 2002; Wouters, Paas, & Van Merriënboer, 2009). This type of measurement requires learners to retrospect only on the task they just finished that is probably still (partly) activated in working memory. The second type of studies asks learners only once at the end of an entire learning or test phase to provide an overall rating (e.g., *single-item scale*: De Koning, Tabbers, Rikers, & Paas, 2010; *NASA-TLX*: Gerjets et al., 2006; Kester et al., 2006; Scheiter et al., 2009; *multiple self-constructed items*: Huk & Ludwigs, 2009). This type of measurement requires learners to provide a retrospective judgment of the cognitive load imposed by a whole sequence of (sub)tasks, which probably needs to be retrieved from long-term memory. Whereas this single retrospective rating is used as an indicator of the cognitive load experienced during the learning phase or test phase, with multiple ratings the experienced cognitive load is obtained by calculating the average of the ratings.

Given these differences in the timing and frequency with which rating scales are applied, there are likely to be differences in the results obtained, and as a consequence, in the conclusions that are drawn. A series of four experiments is presented here, in which it is investigated how the timing and frequency of single-item mental effort measures (Paas scale), that is, after every task in a series or once at the end of a series of tasks, affect the results obtained.

At least two theoretical reasons for favoring the measurement of mental effort immediately after each task over a single measurement at the end of a series of tasks can be given. First, the goal of studies in which effort is measured is often to compare different approaches to instruction, and instructional packages may contain many different types of tasks (e.g., worked examples and problems). When measuring effort only once at the end of the instructional phase, it is impossible to determine to what extent different components of the package contributed to cognitive load during learning. For instance, example study often requires less effort than problem solving (e.g., Paas, 1992; Van Gog et al., 2006), but this information would be unavailable when only a single rating at the end is given. Second, it is clearer what a single measurement after one task reflects than what a single measurement at the end of the series of tasks reflects. That is, participants can be expected to be able to introspect on the effort they invested in a task they just finished, that is likely to still be (partly) activated in working memory. However, when mental effort is measured only once at the end of a series of tasks, the rating will probably also involve information retrieved from long-term memory, and it is unclear whether participants estimate their invested mental effort as an average over all tasks, the last tasks they worked on (i.e., recency effect), the most complex tasks they worked on, or any combination of those possibilities. Experiment 1 therefore implements two conditions, rating after every task vs. rating once at the end of a sequence of tasks, using different task sequences in each condition that contain both simple problems (i.e., low in intrinsic load) and complex problems (i.e., high in intrinsic load). This will allow us to investigate both the differences (if any) between results obtained when measuring mental effort repeatedly or once at the end of a

sequence of tasks, and will provide insight into what a single rating at the end of a sequence might reflect (i.e., does it correspond to the average effort, the effort on the last tasks, or the effort on the most complex tasks indicated in the other condition?).

Experiment 1

Method

Participants

Eighty-seven secondary education students volunteered to participate in this study (39 male; age: $M = 15.60$, $SD = .95$).

Materials

Problems. Six problems were used varying in intrinsic load (i.e., complexity). Three problems were high in intrinsic load (complex), modelled after the problem used by Sweller (1993, p. 6): “Suppose 5 days after the day before yesterday is Friday. What day of the week is tomorrow?” Three problems were low in intrinsic load (simple), for example: “Suppose yesterday was Tuesday. What day of the week is tomorrow?”

Mental effort rating scale. Invested mental effort was measured using the 9-point subjective rating scale developed by Paas (1992), which ranges from (1) very, very low mental effort to (9) very, very high mental effort.

Design and Procedure

Participants were randomly assigned to one of two conditions. In the first (‘Multiple Ratings’; $n = 44$), participants were asked to rate the mental effort they invested in solving each problem directly after completing it (i.e., six times). In the second (Single Rating; $n = 43$), participants were asked to rate the mental effort they invested in solving the problems after completing all problems (i.e., once). The order of the problems was varied within each condition, creating four different sequences, in order to ensure that possible effects would not be due to one particular sequence of tasks we used and the use of different sequences also allowed us to find out

whether single ratings at the end would be affected by recency effects (in which case ending with a simple task and especially a number of simpler tasks should lead to a lower rating than ending with a complex task). Participants were randomly assigned to one of the four sequences: [1] simple-simple-simple-complex-complex-complex (SSSCCC; n = 20), [2] simple-complex-simple-complex-simple-complex (SCSCSC; n = 23), [3] complex-simple-complex-simple-complex-simple (CSCSCS; n = 22), and [4] complex-complex-complex-simple-simple-simple (CCCSSS; n = 22).

Participants were given 1 minute to find the answer to each problem, and they had to do it mentally, that is, they were not allowed to use paper and pen while working out the answer. Depending on assigned condition, they rated their perceived mental effort after each task, or after all six tasks on the nine-point mental effort rating scale. Each problem and each mental effort rating scale was printed on a separate page of A4 paper. Participants were not allowed to proceed to the next problem until the experiment leader signalled that the minute allotted for completing the problem was over.

Results

Performance on each problem was rated as either correct (1 point per problem) or incorrect (0 points). Mean performance on the three simple and the three complex tasks was calculated (i.e., scores were summed and divided by 3; max. mean score = 1). As one would expect, mean performance on the simple tasks was high (M = .93, SD = .20), whereas mean performance on the complex tasks was low (M = .23, SD = .21).

The mean mental effort invested in all six problems in the Multiple Ratings condition was 3.75 (SD = 1.06), whereas that in the Single Rating condition was 5.23 (SD = 1.40). This difference is significant, $t(85) = -5.56$, $p < .001$ (two-tailed), Cohen's d = 1.19.

An ANOVA on effort in the Single Rating condition with sequence as a factor showed no effect of sequence $F(3, 39) = 1.56$, $p > .20$.

A closer look at the data from the Multiple Ratings condition, shows that the mean mental effort invested in the simple problems was 2.03 (*SD* = .79) compared to 5.47 (*SD* = 1.64) in the complex problems. Note that the mean effort invested in the complex problems is almost equal to the mental effort investment indicated by Single Rating condition. A repeated measures analysis with mean mental effort invested in the simple and the complex problems as within subjects variable and sequence as between-subjects variable shows that this difference is highly significant, $F(1, 40) = 227.50$, $MSE = 1.14$, $p < .001$, $\eta_p^2 = .85$. There was no interaction with sequence $F(3, 40) < 1$.

Discussion

This experiment provided some insight into how timing and frequency of measurement affects mental effort ratings. When asked to rate invested mental effort only once after completing all problems in task sequences that contain both simple and complex problems, this rating is higher than the average of ratings over all six tasks. The data suggest that students may base their overall rating on the most complex problems they worked on. As there was no effect of sequence, this seems to be the case irrespective of whether participants worked on the complex problems first (CS) or last (SC) and irrespective of whether tasks were presented in a blocked (SSSCCC and CCCSSS) or alternating (SCSCSC and CSCSCS) fashion.

Because a between-subjects design was used in this study, a possible explanation for these findings might be that participants in the condition that provided a single rating at the end of the series of tasks did not have a good recollection of the tasks or the effort they invested in each task, and mainly remembered the most complex tasks on which they then based their judgment.

Experiment 2

To investigate whether the findings from Experiment 1 can be explained by students not having a good memory representation of the amount of effort they invested in each problem, Experiment 2 uses a within-subjects design, asking participants to rate their invested mental effort

both immediately after each task as well as at the end of the sequence. Rating mental effort immediately after each task may result in better memory of the amount of effort invested in each task, which would be expected to affect the overall effort rating at the end of the sequence.

Method

Participants

Thirty-nine secondary education students volunteered to participate in this study (29 male; age: $M = 16.00$, $SD = .94$).

Materials

Tasks and mental effort rating scales were identical to those used in Experiment 1.

Design and Procedure

Rating was now a within-subjects variable: All participants rated their perceived mental effort immediately after each task, as well as after all six tasks on the nine-point mental effort rating scale. So, there was now only one condition, but as in Experiment 1, four different task sequences were used to which participants were randomly assigned: [1] SSSCCC ($n = 9$), [2] SCSCSC ($n = 10$), [3] CSCSCS ($n = 10$), and [4] CCCSSS ($n = 10$). The procedure was the same as in Experiment 1.

Results

Performance on each problem was again rated as either correct (1 point per problem) or incorrect (0 points). Mean performance on the three simple and the three complex tasks was calculated (i.e., scores were summed and divided by 3; max. mean score = 1). As in Experiment 1, mean performance on the simple tasks was high ($M = .94$, $SD = .17$) and mean performance on the complex tasks was low ($M = .27$, $SD = .26$).

The mean mental effort of ratings on each of the six problems was 3.99 ($SD = 0.93$). The mean effort score provided after the sequence of tasks was again higher than the average over six

tasks: 4.87 (*SD* = 1.43). A repeated measures analysis shows this difference to be significant: $F(1,35)= 39.99$, $p < .001$, $\eta_p^2 = .50$. There was no effect of sequence.

A closer look at the ratings after each task, shows that the mean mental effort invested in the simple problems was 1.92 (*SD* = 0.85) and in the complex problems 6.07 (*SD* = 1.40). Note that the effort rating provided at the end is not as high as the mean of ratings on the complex problems, and a repeated measures analysis on the mean mental effort invested in the three complex tasks and the overall effort rating shows this difference to be significant, $F(1, 35)= 113.36$, $p < .001$, $\eta_p^2 = .75$. There was no effect of sequence.

Discussion

In line with the findings reported in Experiment 1, Experiment 2 also showed that a mental effort rating provided at the end of a sequence of tasks was higher than the mean of the ratings provided after each of the six tasks. Interestingly, however, this overall rating was not as high as the mean rating on the complex problems. Therefore, it is possible that rating invested mental effort after each task indeed led to a somewhat better memory representation, reducing the need to rely on the memory of the most complex tasks. Still, however, the rating at the end of the sequence was significantly higher than the mean of ratings after each task. A potential explanation might be that participants assign the complex problems more weight than the simple problems when asked to give a rating after the whole sequence of tasks.

So, for task sequences that contain both simple and complex problems, the findings from Experiments 1 and 2 imply that researchers measuring cognitive load either after each task or after a sequence of tasks will come to different conclusions regarding participants' mental effort investment in completing the tasks, with the single measure leading to a higher score. An interesting question is what would happen when the task sequence does not contain both simple and complex problems, but tasks of the same complexity level. If students indeed give the complex problems more weight when providing a mental effort rating at the end of a sequence of tasks, the

effect that this leads to a higher score than the average of ratings immediately after each task, should disappear when tasks are all simple or all complex. This question is addressed in Experiment 3.

Experiment 3

Given that tasks of the same level of complexity are used in this experiment, we expect there to be no difference between the average of six ratings obtained after each task and the single rating at the end.

Method

Participants

Forty-five secondary education students volunteered to participate in this study (16 male; age: $M = 15.58$, $SD = 1.22$).

Materials

Tasks and mental effort rating scales were similar to those used in Experiment 1; additional problems were developed to create sequences of six complex (high intrinsic load) and six simple (low intrinsic load) problems.

Design and Procedure

As in Experiment 2, a within-subjects design was used. Participants were randomly assigned to one of two conditions: [1] simple tasks only ($n = 22$), [2] complex tasks only ($n = 23$). All participants rated their perceived mental effort immediately after each task, as well as after all six tasks on the nine-point mental effort rating scale. The rest of the procedure was identical to Experiments 1 and 2.

Results

Performance on each problem was rated as either correct (1 point per problem) or incorrect (0 points). Mean performance on the tasks was calculated (i.e., scores were summed and divided by

6; max. mean score = 1). Performance in the simple tasks conditions was high (M = .99, SD = .04) and performance in the complex tasks conditions was low (M = .34, SD = .18).

A repeated measures analysis, comparing the mean of ratings on each of the six problems to the rating at the end of the sequence, with task complexity as between-subjects factor, shows a highly significant interaction effect, $F(1, 43)= 14.76$, $p < .001$, $\eta_p^2 = .26$. This interaction effect signifies that for the simple problems condition there is no significant difference between the mean of ratings on all six tasks (M = 1.25, SD = .44) and the mean rating at the end of the sequence of tasks (M = 1.14, SD = .35), whereas for the complex problems condition the rating at the end of the sequence of tasks (M = 6.00, SD = 2.09) was higher than the mean of ratings on the six tasks (M = 5.46, SD = 2.00).

Discussion

This experiment showed that in line with our expectation, there was no difference between the mean of ratings provided immediately after each task and the rating at the end when tasks were all low in intrinsic load. In contrast to our expectation, however, this was not the case when tasks were all high in intrinsic load. On the complex (high intrinsic load) problems a similar pattern emerged to that found in Experiments 1 and 2, with the rating at the end being higher than the average of ratings provided immediately after each task.

This suggests that the findings of Experiments 1 and 2, in which task sequences contained both simple and complex problems, cannot be explained by students assigning more weight to the complex problems when providing a mental effort rating at the end of the sequence of tasks. A possible explanation for these results is one that was previously raised (Experiment 2), that is, the memory representations (or lack thereof) of the tasks and mental effort invested in those tasks. Participants' limited working memory resources have to be distributed between task performance and monitoring activities (cf. Kanfer & Ackerman, 1989), which is not too difficult on simple tasks, but when they are not aware that monitoring is required, learners will probably devote all available

resources to trying to find a solution on the complex tasks (see also Yeo, Loft, Xiao, & Kiewitz, 2009). As a consequence, participants are not likely to build an adequate memory representation of those tasks. However, if participants would know beforehand that they will be required to rate mental effort afterwards, they may devote some cognitive capacity to monitoring, which should result in a more accurate memory representation, which should be reflected in the mental effort rating.

Experiment 4

In this experiment, participants were informed beforehand that they would have to retrospectively introspect on the mental effort they invested in solving the problems. Next to two conditions in which a single rating at the end is required (knowing or not knowing this), we also added conditions in which retrospective ratings have to be provided for each of the six problems (i.e., providing a mental effort rating per problem, after having completed all problems), knowing or not knowing this. The latter conditions were added to be able to investigate the effects of timing of ratings as well as of monitoring more thoroughly. That is, regarding timing, these conditions allow for investigating whether there are differences between multiple ratings and a single rating when all are done retrospectively. In addition, it allows detecting effects of monitoring on a more detailed level: when knowing a rating is required, participants should be better able to remember the order of the simple and complex problems, and thus more accurately rate the effort they invested in each problem than when not knowing this.

It is expected that participants' knowing that they have to retrospectively rate mental effort will lead to more active performance monitoring and hence, a difference in ratings between the conditions. Specifically, given the findings from the previous experiments, one would expect this difference to be that the mental effort ratings are lower in the conditions in which participants knew they would be asked to rate their perceived mental effort retrospectively than in the conditions in which they did not know this.

Method

Participants

Participants were 107 secondary education students (51 male; age: M = 16.46, SD = .82).

Materials

Tasks and mental effort rating scales were identical to those used in Experiments 1 and 2. Participants who were asked to retrospectively rate mental effort per task, were given a list saying task 1, task 2, task 3, et cetera. That is, no clue was given as to the content of each task. In the conditions in which participants were informed beforehand that they would have to rate their invested mental effort after completing all tasks, they were not informed whether they would have to do this per task or in a single rating; they were only told: “Afterwards, you will be asked to indicate how much mental effort you invested in solving those problems”.

Design and Procedure

Participants were randomly assigned to one of four conditions: [1] giving a single mental effort rating after completing all six problems, knowing that this would be requested (Single Rating –Knowing; n = 28), [2] giving a single mental effort rating after completing all six problems, not knowing that this would be requested (Single Rating –Not Knowing; n = 27), [3] giving a mental effort rating on each of the six problems after completing all six, knowing that this would be requested (Multiple Ratings –Knowing; n = 26), and [4] giving a mental effort rating on each of the six problems after completing all six, not knowing that this would be requested (Multiple Ratings; Not Knowing; n = 26). As in Experiments 1 and 2, four sequences of tasks were used to which participants in each condition were randomly assigned. All participants retrospectively rated their perceived mental effort (i.e., after completing all six tasks) on the nine-point mental effort rating scale, either providing one overall rating or a rating per task. The rest of the procedure was identical to that of Experiments 1, 2, and 3.

Results

Performance on each problem was rated as either correct (1 point per problem) or incorrect (0 points). Mean performance on the three simple and the three complex tasks was calculated (i.e., scores were summed and divided by 3; max. mean score = 1). In line with Experiments 1, 2, and 3, performance on the simple tasks was high (M = .95, SD = .18) and performance on the complex tasks was low (M = .23, SD = .25).

An ANOVA on the mean mental effort ratings given in the Multiple Ratings conditions shows a significant difference, $F(1, 50) = 5.30$, $p < .05$, $\eta_p^2 = .096$. In line with our expectation, the mean mental effort was lower in the Knowing condition (M = 4.08, SD = 1.17) than in the Not Knowing condition (M = 4.78, SD = 1.01).

An ANOVA on the mental effort ratings given in the Single Rating conditions also shows a significant difference, $F(1, 53) = 7.03$, $p < .05$, $\eta_p^2 = .117$. In line with our expectation, mental effort was lower in the Knowing condition (M = 4.68, SD = 1.18) than in the Not Knowing condition (M = 5.59, SD = 1.37).

However, again there seems to be a large difference between providing a single rating and multiple ratings even when those are also provided retrospectively, with the single rating again seeming to be higher. An ANOVA comparing the mean of the six ratings with the single rating between conditions, shows that this difference is significant, $F(1, 105) = 8.40$, $p = .005$, $\eta_p^2 = .074$, with the single rating being higher (M = 5.13, SD = 1.35) than the mean of the six ratings (M = 4.43, SD = 1.14).

A closer look at the mental effort rated retrospectively per task for the simple and complex problems, shows that the mean mental effort invested in the simple problems in the Knowing condition is 2.28 (SD = 1.63), but in the Not Knowing condition it was 3.67 (SD = 1.88); this difference is significant, $F(1, 50) = 8.04$, $p < .01$, $\eta_p^2 = .139$. On the complex problems, there was no significant difference between Knowing (M = 5.87, SD = 1.76) and Not Knowing (M = 5.88, SD = 1.98), $F(1, 50) < 1$.

Discussion

In line with our expectation, Experiment 4 shows that knowing that a retrospective mental effort rating is required, will alter (i.e., lower) those retrospective ratings. This suggests that the performance monitoring explanation might indeed explain the findings from Experiments 1, 2, and 3, which found single retrospective ratings to be higher than the mean of ratings over six tasks. It seems that telling participants beforehand that a retrospective mental effort rating will be required, lowers this rating. However, this does not necessarily make it more in line with the average of ratings over six tasks, as a comparison (by visual inspection) with the data from Experiment 1 shows (Experiment 2 is not considered here as it applied a within-subjects design, whereas both Experiment 1 and 4 used a between-subjects design and are as such more comparable). In Experiment 1, the mean of six ratings provided immediately after each task was 3.75, and of the single rating at the end was 5.23. In Experiment 4, the group that had to provide a single rating not knowing this, had a mean of 5.59, whereas the condition that knew a rating would be required, had a mean of 4.68. So although the ratings by both groups who provided a single rating not knowing this would be required, seem quite equal, and the rating from the group who was aware that a rating would be required is lower, this rating (4.68) still seems substantially higher than the average over six tasks (3.75).

Interestingly, the mean of six ratings in the condition in which a retrospective rating had to be provided for each task and participants were aware of this (i.e., 4.08), is still somewhat higher, but comes much closer to the mean of six ratings provided immediately after every task (i.e., 3.75) than any of the other conditions. This suggests that when participants are instructed that they will have to provide a retrospective rating of their mental effort, they may be quite accurate in providing ratings per task. This again suggests that they are monitoring more actively, which is further supported by the analysis of the differences in complex and simple task ratings between participants who had to provide a retrospective mental effort rating per task, knowing or not

knowing that mental effort rating would be required. There was a difference in the ratings provided on the simple tasks, with those in the condition that knew being lower (i.e., 2.28, which does not deviate much from the 2.03 found in Experiment 1) than in the condition that did not know (i.e., 3.67), which probably means that learners in the condition that knew they would be asked to rate mental effort remembered pretty well where the simple tasks were located in their sequence, whereas participants who did not know this beforehand were probably not very sure about the order of the tasks, and therefore gave higher ratings on the simple problems as well.

General Discussion

The findings from Experiment 4 highlight the important role of monitoring when rating retrospectively. The average of retrospectively provided ratings per task comes closest to the average of ratings provided immediately after each task in Experiment 1. However, even when all ratings are done retrospectively, a single rating resulted in a higher score than the average of six ratings, which suggests that monitoring cannot fully explain the difference between a single retrospective rating at the end of a sequence of tasks and the average of multiple ratings. Rather, taken together, the findings from the four experiments presented here suggest that possibly, participants are not considering the effort invested in every single task when providing a single mental effort rating at the end of a series of tasks, but instead, perceive the series as one task to be rated. If so, the number of tasks in the series and/or duration of the series might be a potential cause of the higher rating at the end, given that time also plays a role in experienced cognitive load (see Xie & Salvendy, 2000). Another potential cause might lie in the so-called positive-negative asymmetry effect, which shows that “In general, …, negative information receives more processing and contributes more strongly to the final impression than does positive information.” (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; p. 323-324). If complex problems are experienced as more negative, they might contribute more to the overall impression than simple problems. This might also explain the findings from Experiment 4, that single retrospective ratings became lower when

participants knew such a rating was required, and that participants who provided retrospective ratings for each task and knew they had to do so seemed to remember better where in the sequence the simple tasks were located. That is, monitoring task complexity may have altered those participants' impression of the task sequence, making the more complex tasks somewhat less dominant in the final impression. Future research should address these explanations and try to uncover what exactly participants base their mental effort ratings on, for example by applying concurrent or (cued) retrospective verbal reporting techniques to ratings of mental effort (Ericsson & Simon, 1993; Van Gog, Paas, Van Merriënboer, & Witte, 2005).

Another interesting question for future research to address, is whether the same timing and frequency effects would apply to other subjective rating scales that are sensitive to detecting variations in task difficulty which are associated with cognitive load, such as perceived difficulty ratings, which have also been applied during tasks (e.g., Ayres, 2006), after each task (e.g., Marcus, Cooper, & Sweller, 1996), or after a sequence of tasks (e.g., Kalyuga, Chandler, Tuovinen, & Sweller, 2001; Pollock, Chandler, & Sweller, 2002). Finally, it would be interesting to investigate in future studies what the relationship is between multiple ratings or single ratings and learning outcomes, both in terms of test performance and in terms of effort investment on the test. A potential limitation of the present study is that we did not use learning tasks or measure learning outcomes (i.e., we assessed effort and direct performance, not future performance), whereas most studies in which mental effort is measured are conducted with the aim of comparing different approaches to instruction. One might argue that it is not so bad if multiple ratings and a single rating do not correspond, as long as both would lead to the same conclusions about differences between instructional conditions. However, as mentioned before, when measuring effort only once at the end of an instructional phase, it is impossible to determine to what extent different components of the instructional package contributed to cognitive load during learning. Moreover, depending on what the single rating is based on exactly, it is not unthinkable that subtle differences

between conditions would be missed when only single ratings at the end of an instructional phase are used.

Until more is known about what mental effort ratings at the end of a series of tasks reflect, the results of the four experiments presented here seem to suggest that measuring mental effort after each task is preferable over a single measure at the end of a series of tasks. One might argue based on the findings from Experiment 4 that providing mental effort ratings retrospectively is not a problem as long as this is done per task, and participants are informed beforehand that such ratings are required. However, this may change when task sequences are longer, or when time on task is longer, because monitoring task content and effort investment might be more difficult on tasks or sequences of longer duration. It might therefore be advisable to apply a rating scale multiple times during a task when it is of relatively long duration (cf. first experiment of Van Merriënboer et al., 2002; Yeo & Neal, 2004, 2008; see also Ayres, 2006, for an application of perceived difficulty ratings during a task). Moreover, repeated ratings over time in a series of tasks or in a task of longer duration will provide information on the fluctuation in cognitive load (i.e., instantaneous load; Xie & Salvendy, 2000) that is lost when only a single rating at the end is obtained.

References

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic load within problems. *Learning and Instruction, 16,* 389-400.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5,* 323-370.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (rev. ed.). Cambridge, MA: The MIT Press.

Gerjets, P., Scheiter, K., & Catrambone, R. (2006). Can learning from modular worked examples be enhanced by providing instructional explanations and prompting self-explanations? *Learning and Instruction, 16*, 104-121.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139-178). Amsterdam: Elsevier Science.

Hoffman, B., & Schraw, G. (2010). Conceptions of efficiency: Applications in learning and problem-solving. *Educational Psychologist, 45*, 1-14.

Huk, T., & Ludwigs, S. (2009). Combining cognitive and affective support in order to promote learning. *Learning and Instruction, 19*, 495-505.

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.

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/ aptitude-treatment approach to skill acquisition. *Journal of Applied Psychology, 74*, 657–690.

Kester, L., Lehnen, C., Van Gerven, P. W. M., & Kirschner, P. A. (2006). Just-in-time, schematic supportive information presentation during cognitive skill acquisition. *Computers in Human Behavior, 22*, 93-112.

Kirschner, F., Paas, F., & Kirschner, P. A. (2009). Individual and group-based learning from complex cognitive tasks: Effects on retention and transfer efficiency. *Computers in Human Behavior, 25*, 306-314.

Kirschner, F., Paas, F., Kirschner, P. A. & Janssen, J. (2011). Differential effects of problemsolving demands on individual and collaborative learning outcomes. *Learning and Instruction, 21*, 587-599.

Marcus, N, Cooper, M. G., & Sweller, J. (1996). Understanding instructions. *Journal of Educational Psychology*, *88*, 49 - 63

Nievelstein, F., Van Gog, T., Van Dijck, G., & Boshuizen, H. P. A. (2011). Instructional support for novice law students: Reducing search processes and explaining concepts in cases. *Applied Cognitive Psychology, 25*, 408–413.

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., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*, 63-71.

Paas, F., Tuovinen, J., Van Merriënboer, J. J. G., & Darabi, A. (2005). A motivational perspective on the relation between mental effort and performance: Optimizing learner involvement in instruction. *Educational Technology, Research & Development, 53*, 25-33.

Paas, F., & Van Merriënboer, J. J. G. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors, 35*, 737–743.

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.

Pollock, E., Chandler, P., & Sweller, J. (2002). Assimilating complex information. *Learning and Instruction, 12*, 61-86.

Scheiter, K., Gerjets, P., Vollmann, B., & Catrambone, R. (2009). The impact of learner characteristics on information utilization strategies, cognitive load experienced, and performance in hypermedia learning. *Learning and Instruction, 19*, 387-401.

Spanjers, I. A. E., Wouters, P., Van Gog, T., & Van Merriënboer, J. J. G. (2011). An expertise reversal effect of segmentation in learning from animated worked-out examples. *Computers in Human Behavior, 27*, 46-52.

Sweller, J. (1993). Some cognitive processes and their consequences for the organisation and presentation of information. *Australian Journal of Psychology, 45*, 1-8.

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*, 251-295.

Van Gerven, P. W. M., Paas, F., Van Merriënboer, J. J. G, & Schmidt, H. G. (2004). Memory load and task-evoked pupillary responses in aging. *Psychophysiology, 41*, 167-175.

Van Gog, T., Kester, L., & Paas, F. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices' learning. *Contemporary Educational Psychology, 36*, 212–218.

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 Gog, T., Paas, F., & Van Merriënboer, J. J. G. (2008). Effects of studying sequences of process-oriented and product-oriented worked examples on troubleshooting transfer efficiency. *Learning and Instruction, 18*, 211-222.

Van Gog, T., Paas, F., Van Merriënboer, J. J. G., & Witte, P. (2005). Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology: Applied, 11*, 237-244.

Van Merriënboer, J. J. G., Schuurman, J. G., De Croock, M. B. M., & Paas, F. (2002). Redirecting learners' attention during training: Effects on cognitive load, transfer test performance and training efficiency. *Learning and Instruction, 12*, 11-37.

Wouters, P., Paas, F., & Van Merriënboer, J. J. G. (2009). Observational learning from animated models: Effects of modality and reflection on transfer. *Contemporary Educational Psychology, 34*, 1-8.

Xie, B., & Salvendy, G. (2000). Review and reappraisal of modeling and predicting mental workload in single-and multitask environments. *Work & Stress, 14*, 74-99.

Yeo, G., Loft, S., Xiao, T., & Kiewitz, C. (2009). Goal orientations and performance: Differential relationships across levels of analysis and as a function of task demands. *Journal of Applied Psychology, 94*, 710–726.

Yeo, G. B., & Neal, A. (2004). A multilevel analysis of effort, practice and performance: effects of ability, conscientiousness, and goal orientation. *Journal of Applied Psychology, 89*, 231–247.

Yeo, G., & Neal, A. (2008). Subjective cognitive effort: A model of states, traits, and time. *Journal of Applied Psychology, 93*, 617–631.