Behav Res (2017) 49:1905–1919
DOI 10.3758/s13428-016-0833-y

Pupillary response to complex interdependent tasks: A cognitive-load theory perspective

Ritayan Mitra[1] · **Karen S. McNeal**[1] · **Howard D. Bondell**[2]

Published online: 7 December 2016

Abstract Pupil dilation is known to indicate cognitive load. In this study, we looked at the average pupillary responses of a cohort of 29 undergraduate students during graphical problem solving. Three questions were asked, based on the same graphical input. The questions were interdependent and comprised multiple steps. We propose a novel way of analyzing pupillometry data for such tasks on the basis of eye fixations, a commonly used eyetracking parameter. We found that pupil diameter increased during the solution process. However, pupil diameter did not always reflect the expected cognitive load. This result was studied within a cognitive-load theory model. Higher-performing students showed evidence of germane load and schema creation, indicating use of the interdependent nature of the tasks to inform their problem-solving process. However, lower-performing students did not recognize the interdependent nature of the tasks and solved each problem independently, which was expressed in a markedly different pupillary response pattern. We discuss the import of our findings for instructional design.

Keywords Task-evolved pupillometry · Cognitive-load theory · Eye tracking · Graph reading

Cognitive load, also referred to as *processing load* or *mental effort*, is a term that has long been used by psychologists to represent the loading of working memory in response to a particular task. Working memory has been found to be extremely limited in human beings in terms of the number of independent elements it can store, their interactivity, and the duration of storage (Baddeley & Hitch, 1974). Traditionally, on-task cognitive load has been measured by empirical measures such as rating scales to differentiate between easy and hard tasks, and with performance data such as the time of response and accuracy (Cierniak, Scheiter, & Gerjets, 2009; Kalyuga, Chandler, Tuovinen, & Sweller, 2001; Leppink, Paas, Van der Vleuten, Van Gog, & Van Merriënboer, 2013; Paas, 1992).

More recently, biophysical proxies have been used for continuous and more precise measurement of cognitive load. Biophysical proxies of cognitive load, such as electroencephalography (EEG) (Antonenko, Paas, Grabner, & van Gog, 2010) and functional magnetic resonance imaging (fMRI; Whelan, 2007) are believed to capture cognitive load effectively by measuring indicators of mental activity such as glucose consumption (PET scanning) or oxygen uptake (fMRI). Among such biophysical measures, pupil diameter has shown much promise in the measurement of cognitive load (Van Gerven, Paas, Van Merriënboer, & Schmidt, 2004). The muscles of the iris work in tandem with the autonomic nervous system to dilate or contract the pupil, usually between diameters of 2 and 7 mm, in response to ambient light conditions. Along with this response to brightness, pupils also contract or dilate in response to emotional and mental stimuli. The principle behind pupillometry is simple—when brightness and the emotional content of a stimulus are kept constant, pupillary response can be used as a proxy for cognitive load.

The earliest recognition of the pupil diameter change to stimuli other than light appears to have been made in the late nineteenth century (Heinrich, 1896; Schiff, 1875). Subsequently, German neurologist Oswald Bumke noticed what was then known as *Bumke syndrome* or *Bumke pupil*:

✉ Ritayan Mitra
rmitra@ncsu.edu

[1] Department of Marine, Earth and Atmospheric Science, North Carolina State University, Raleigh, North Carolina, USA

[2] Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA

pupillary dilation in response to anxiety or other psychic stimuli (cited in Hess, 1972). Not until Hess and Polt's work on pupillary response to emotional states, such as affection, interest, or sexual arousal (Hess & Polt, 1960), or mental activity (Hess & Polt, 1964) did the topic receive widespread attention in the cognition community. Closely following Hess and Polt's seminal work, the comprehensive body of work by Kahneman and Beatty (Beatty & Kahneman, 1966; Kahneman & Beatty, 1966) laid the foundation for pupillometry studies directed at measuring the cognitive loads from a variety of tasks.

Ever since, pupillometry has been routinely used in psychophysiological studies of cognitive load that have looked at the following four primary categories of tasks (in terms of the mental attribute each task is intended to represent): memorizing digit spans (memory; Granholm, Asarnow, Sarkin, & Dykes, 1996; Peavler, 1974; Piquado, Isaacowitz, & Wingfield, 2010), word processing (language skills; Ben-Nun, 1986; Just & Carpenter, 1993; Hyönä, Tommola, & Alaja, 1995; Schmidtke, 2014), multiplication (mental math skills; Ahern & Beatty, 1979; Landgraf, Van der Meer, & Krueger, 2010), and search (perception; Porter, Troscianko, & Gilchrist, 2007). Modern studies continue to bolster these categories of tasks, which were first outlined by Beatty (1982).

It is worth noting that the tasks used in traditional pupillometry experiments consisted of simple steps. For example, memorizing a sequence of digits is a relatively simple operation (although the difficulty level varies with the length of the sequence). Similarly, the task of multiplying numbers can also be considered simple. Such studies use a methodology that is ideal for assessing the working memory load associated with simple tasks. However, the methodology is not suitable when we want to assess the cognitive load associated with more complex tasks. Such tasks invariably require processing of information from multiple sub-tasks. For example, if the task involves map reading, then the subtasks would be looking up the legend, identifying symbols, and searching for those items on the map If the cognitive-load variation during map reading is to be ascertained, these individual components need to be identified in the pupillometry data, so that pupil data can be averaged over meaningful cognitive epochs of the task. In other words, to evaluate the cognitive load arising from a complex task, we first have to identify the subtasks that are relevant, and averaging of pupil data must be carried out only within these subtasks for any meaningful interpretation. This requirement poses a challenge, because pupillometry data do not come with locational information (in the example above, we must know whether the individual was focusing on the legend or the map). However, an eyetracker also captures gaze data that can be used for this purpose. There appears to be a disciplinary chasm between researchers who are interested in pupil dilation versus gaze data. The former branch of study is largely referred to as *pupillometry*, and the latter as *eyetracking*. The two lines of inquiry seem to have developed independently, although both types of data are captured simultaneously by the eyetracker. One of the methodological advancements we propose in this study involves using gaze data to identify epochs or subtasks within a complex task, and using that knowledge to delineate pupil data for a meaningful interpretation of the cognitive load associated with a complex task.

Another issue with pupillometry experiments comprising simple tasks arises from the relatively independent natures of these tasks. For example, in a digit span task, the participant is expected to encounter increasingly longer digit strings that must be recalled. None of the tasks informs or scaffolds subsequent tasks; no skill development or expertise acquisition is either required or encouraged. In other words, each task is independent within the experimental paradigm. However, educational contexts are cumulative (interdependent)—that is, solving a problem informs one's future approach to solving problems of a similar nature. Unless a methodology is developed to assess cognitive load between such interdependent sets of tasks—that is, each task building upon previous tasks—important educational parameters such as skill development and knowledge transfer cannot be assessed. Pupillometry profiles from such interdependent tasks can potentially monitor such important educational parameters.

In this study, we develop a methodology for using a combination of eyetracking and pupillometry data for a set of tasks that are complex and interdependent. We discuss the results within a cognitive-load theory (CLT) framework and demonstrate the potential of pupillometry in education research.

Method

Participants

Participants were recruited from a large southeastern U.S. university through the distribution of flyers outside high-enrollment classrooms, and also through announcements in introductory courses. The 29 participants (15 women and 14 men) included 15 freshmen, 12 sophomores, one junior, and one senior student, ranging in age from 18 to 22 years (M = 19.1 years). The majors (if declared) included engineering, management or business administration, environmental science, and earth and atmospheric sciences. All participants had normal or corrected-to-normal eyesight. One participant quit the experiment midway, and another was excluded due to excessive blinking. The participants were compensated with $10 Amazon gift cards. The study was approved by the local institutional review board, and all participants gave informed written consent.

Study design

The study included three questions of successively increasing complexity, based on a single graphical input (Fig. 1). The line graph plotted the coffee production of a region over a period of 1 year. Months were plotted on the *x*-axis, and number of bags was plotted on the *y*-axis. The graph also included a conversion factor that could be used to translate production from bags to pounds. Figure 1 shows the three questions. The first question (Q1) did not involve any mathematical calculation. The second and third questions (Q2 and Q3) required simple addition and multiplication. Question 1 can be considered independent of Q2 and Q3, whereas Q2 and Q3 were dependent. Question 2 required understanding of the conversion factor and performing one mental calculation, whereas Q3 required three such calculations and an additional summing operation or, alternatively, three summing and one conversion operation.

Procedure

The participants read directions from a sheet of paper on which they were briefed on the nature of the task. They were told that they would be asked three questions based on the same chart. It was explained that there would be enough time for them to answer the questions and that the first question would be relatively simpler than the next two, which would involve some mental calculations. Throughout the experiment they would not have to click or press any keys, since the slides would change after a precise time interval. The participants were asked to verbally communicate the answer only once and were asked not to change their answers once spoken aloud. They were asked not to verbalize the mathematical steps they were performing or to look away from the screen during the experiment.

The experiment consisted of eight slides in the following order (at the given times): white slide (5 s), black (5 s), gray (5 s), Q1 (30 s), gray (5 s), Q2 (35 s), gray (5 s), and Q3 (45 s; Fig. 1). The white and black slides were used to normalize the pupil data (Piquado et al., 2010), and the gray slides were used to provide a neutral start point for the pupillometry data for each question. The time for each question was decided through a pilot study, such that there was enough time for the questions to be answered. However, the participants were unaware of the time they would have to answer each question. Instead, they were assured that the time would be sufficient to answer each question. The first question was the simplest, but the slide time was fixed at 30 s to familiarize the participants with the chart content, and also to give them confidence that the time provided would be sufficient. This ensured minimization of any unintended performance anxiety arising from time constraint.

The participants were seated comfortably in front of a Tobii TX300 Eye Tracker (23-in. screen with 1,920 × 1,080 pixel resolution). This system captures data at 300 Hz (or every 3.3 ms). The height of the chair was adjusted for each participant during calibration. The participants were seated at a distance of 60–75 cm from the screen. The participants'

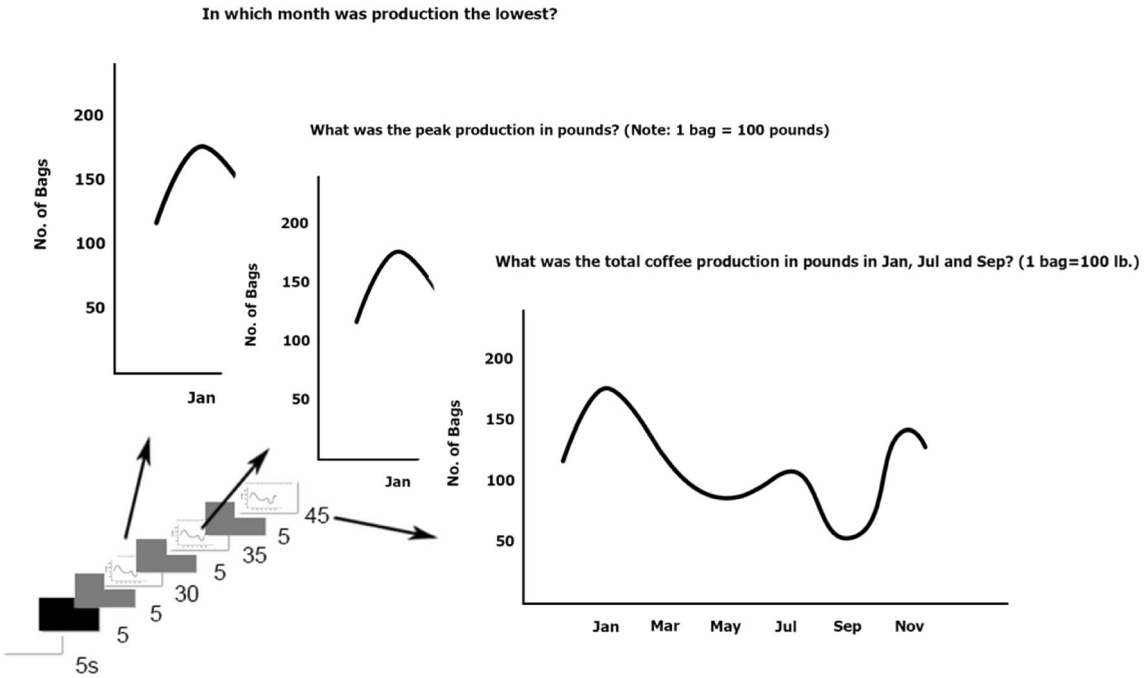

Fig. 1 Experimental design and question slides. *Bottom left* inset shows the slide progression, starting with a white slide at the beginning. The allocated time (in seconds) for each slide is shown next to the slides. The three question slides are magnified

heads were not restrained, because Tobii systems can automatically compensate for small natural head movements (e.g., at 65 cm the participant can move his head within a vertical area of 37 × 17 cm at a maximum speed of 50 cm/s without losing tracking sensitivity). A secondary scene camera was used to record videos of the participants. Both audio (when the answer was spoken) and lip movements are useful in the accurate manual logging of the answer time required for pupillometry data analysis.

Data analysis

Cognitive framework Pupillometry studies have usually been conducted to identify task-evolved pupillometry response (TEPR) related to simple tasks such as searching or remembering a sequence of numbers. In contrast, problem solving with graphical information is a complex task that requires reading, comprehension, and mathematical calculations. Therefore, one goal of the present study was to broaden the scope of pupillometry applications by providing a blueprint for the analysis of such compound tasks. Consequently, we were confronted with the issue of identifying the cognitive steps that go into graphical problem solving. A vast literature deals with aspects of graphing that can facilitate chart interpretation (see Pinker, 1990, and Shah & Hoeffner, 2002, for an introduction and review, respectively). This study, however, is less concerned with the comprehensibility of graphs as it pertains to formatting and readability. Instead, here we seek to understand the cognitive processes behind graphical problem solving. As such, even though we acknowledge that both aesthetics and background information of participants will affect performance, our goal was to study the cognitive processes behind graphical problem solving, and not aspects of problem design. Therefore, we will not focus on the elements of graphing in this study, and assume that the interpretability of the simple graph used for the experiment is uniform for the population.

Given the premise above, we identified three key steps that were needed to answer the questions posed in our study. The principal idea was to keep the cognitive framework broad enough to accommodate variation among the participants, but distinct enough to aid in the discrimination of cognitive processes. Consequently, we identified three major subtasks that can be associated with such problem solving: question comprehension (QC), information look-up (IL), and mental calculation (MC). QC pertains to understanding the question, IL involves looking up and caching the information necessary for problem solving, and MC involves using the stored information (performing calculations) to answer the question.

A post-hoc analysis of the eyetracking data confirmed the suitability of this framework. Figure 2 demonstrates a typical gaze pattern for Q2. The total time of response of this participant was 32 s. In the first 4 s, the gaze pattern indicates the participant was reading the question (QC). Between 4 and 15 s, the gaze pattern showed maximum spread across the slide. The eyes fixated on the *y*-axis values before moving on to the *y*-axis label. Subsequently, the gaze pattern indicates a brief reading of the *x*-axis values before returning to the middle of the figure. After revisiting some of the previous locations, the last fixations in this segment were near the values 200 and 150 on the *y*-axis, a key location for answering Q2. This segment was characterized by broad sweeps, discarding unnecessary information (months), and revisiting locations (e.g., 200 and 150 on the *y*-axis) that contained information necessary to solve the question (IL). The next segment is a very short span, between 15 and 17 s, during which the participant revisited the question (QC). The following segment was characterized by the longest period of focused attention, between 17 and 27 s, as is evident from the longer fixation durations (sizes of the circles), as well as from the fewer locations visited. Attention seems to be focused only on the values on the *y*-axis and the conversion factor at the end of the question, the only two pieces of information necessary to answer the question. This indicates that this segment was dominated by deep thought in which MC took place. Between 27 and 29 s, the gaze quickly moved over the question (QC), indicating a confirmatory action before the participant answered the question. Between 29 and 32 s, the gaze pattern indicates another confirmatory (MC) move in which the attention refocused on the areas visited during MC. The above pattern can be coded as QC–IL–QC–MC–QC–MC. A more focused pattern could be simply QC–IL–MC, and a less sure pattern would look like QC–IL–QC–IL–MC–QC–MC. Although it is difficult to identify when exactly a participant transitioned from one step to the next—or whether some parallel processing occurred at transitions or on a broader scale, despite studies that have shown the dominance of serial over parallel information processing (Körner, 2011; Körner, Höfler, Tröbinger, & Gilchrist, 2014)—the scheme above is expected to capture the major steps of the problem-solving process.

Using the scheme above, and under the following assumptions, we can use TEPR in a novel way to decipher the cognitive load associated with compound tasks. First, we assumed that each response could be divided into some combination of QC, IL, and MC only. Therefore, unconstrained thoughts, which might affect pupil diameter before answering, are not included in this scheme. However, any of the components can be excluded in a response. For example, Q1 would have no MC component, and on-task question exploration within a limited duration would be expected to have negligible unconstrained thoughts. Second, we assumed that the total response time would have more QC

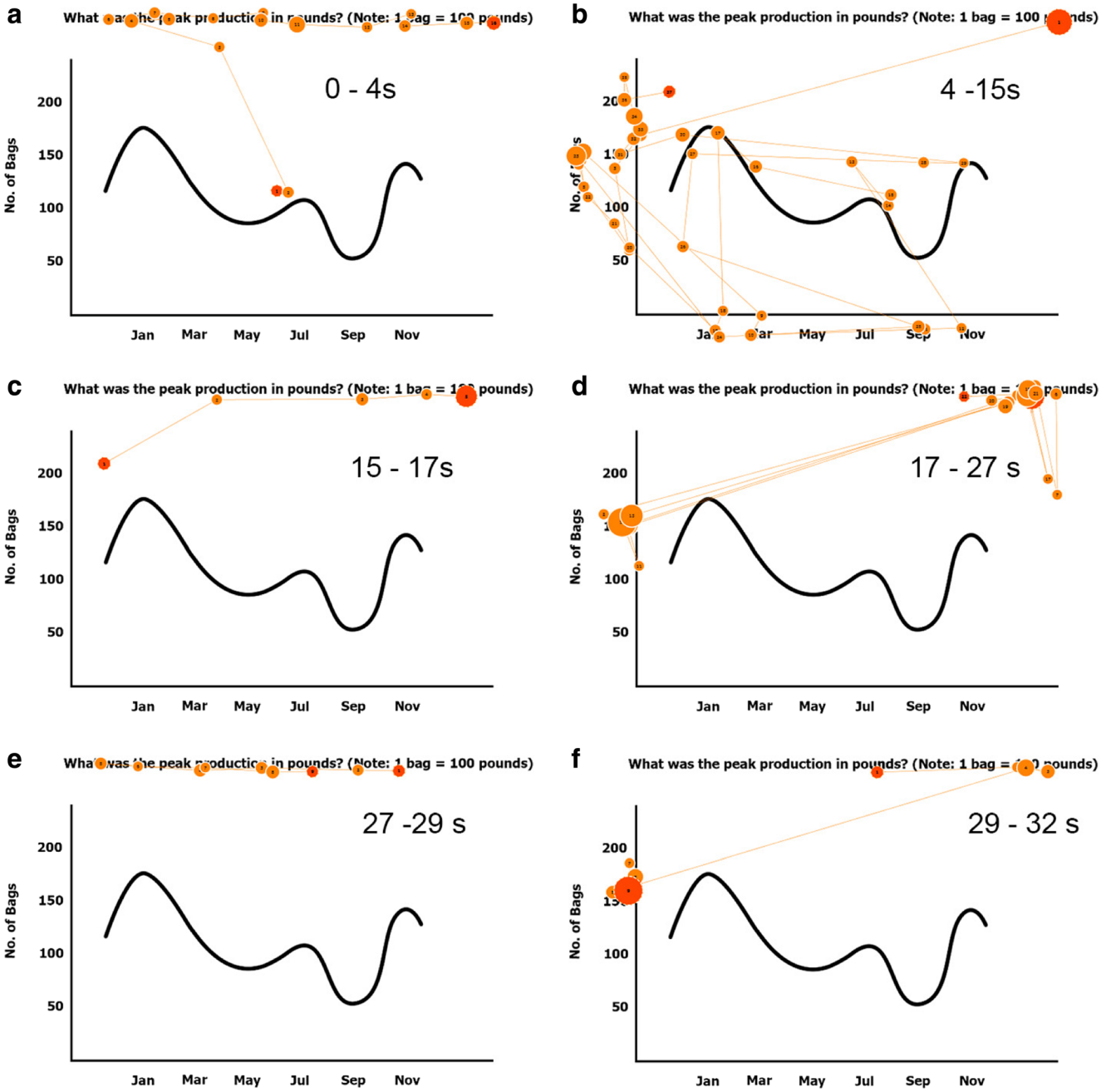

Fig. 2 Time evolution of fixations. The circles indicate momentary halts in the gaze pattern (fixations), and the lines show the gaze transiting between fixations (saccades). Fixations indicate attention, and saccades indicate the intermediate time spent between two points of attention. The eyetracker captures the *x*–*y* coordinates of the gaze, and a filter assigns fixation or saccadic characteristics to those coordinates. We used the I-VT algorithm (Komogortsev et al. 2010; available within Tobii Studio 3.3) to classify the gaze locations into fixations and saccades. Each figure (**a**–**f**) starts with the last fixation in the previous slide. *Deep-shaded circles* indicate the first and last fixations within a particular time slice. The cognitive steps from panels a to f are QC, IL, QC, MC, QC, and MC, respectively

and IL in the first half and more MC during the second half, so that TEPR would largely reflect QC and IL in the first and MC in the second half. If MC is absent, then the first half of the TEPR would be dominated by QC and the second half by IL. The second assumption is further supported by Fig. 2, in which the first 17 s were spent on QC and IL, and the remaining time was spent largely on MC.

Grading scheme Question 1 simply required comprehension of the question and looking up a single value on the *x*-axis. However, Q2 required comprehension of the question, identifying the need to use the conversion factor (from bags to pound), looking up of a single value on the *y*-axis, and performing a single multiplication. Question 3 required all that was necessary for the second question, but there were

three additional calculations to be performed (either three multiplication and one addition or three addition and one multiplication). The grading scheme was designed to differentiate between successful and unsuccessful engagement of these cognitive steps. The questions carried points proportional to the difficulty level, and the grading process involved step-marking (Q1, 1 point; Q2, 3 points; Q3, 5 points; maximum score = 9 points).

Since the first question (whose correct answer was September) only required understanding the question and looking up the month, it was scored as 1 point. The second question (whose correct answer was 17,500), on the other hand, required three separate steps. One point was awarded for basic understanding of the question; any answer between 150 and 200 would indicate fulfillment of this criterion, because an answer in this range would indicate that the participant had successfully interpreted the word "peak," regardless of the accuracy of the answer. Another point was linked to answers that were of any order higher than 100. For example, 17,500 or 1,750 would earn one more point, because they would signify understanding the difference between bags and pounds and an effort at using the conversion factor. Finally, 1 point was awarded for accuracy when a participant stated the correct answer of 17,500. Points for accuracy were deducted only if the mistake was independent of those arising from other deductions. Therefore, for an answer of 175,000, only 1 point would be deducted, either for accuracy or for not getting the order correct. Question 3 (whose correct answer was 32,500) was structured in a similar way. Three points were awarded if the first two digits of the number was close to 32, as in 30 or 34, showing that the effort included some attempt at conversion. That was because this required understanding the question and addition of the values for the respective months. Therefore, even an answer of 3,200 would score 3, but not 320, because this answer indicated not understanding the difference between bags and pounds. Any answer that showed a miscalculation during the addition step of the three values would qualify for a 1-point deduction. An answer of 320 would qualify for a 2-point deduction. Any other response would qualify for a 3-point deduction. One point was awarded for getting the right order of magnitude (i.e., thousands). Finally, 1 point was awarded for accuracy, and this was deducted only when the mistake was independent. For example, an answer of 34,000 or 29,000 to Q3 would qualify for this 1-point deduction.

Pupil data The Tobii TX300 operating at its peak frequency generates one measurement every 3.33 ms. To systematically study the vast amount of data generated during 130 s of recording per person (contributing to more than 1.3 million records), a standalone software, Pupilreduce 1.0, was developed in-house. All processing of the pupillometry data was carried out with this software. The raw data were first cleaned of blinks and gaps (Fig. 3). Any missing data were identified as gaps. Blinks were identified by a moving window of size 50 ms. Any value less than three standard deviations of the mean of the measurements within that window qualified as a blink and was discarded. All missing values (gaps and blinks that were removed) were then linearly

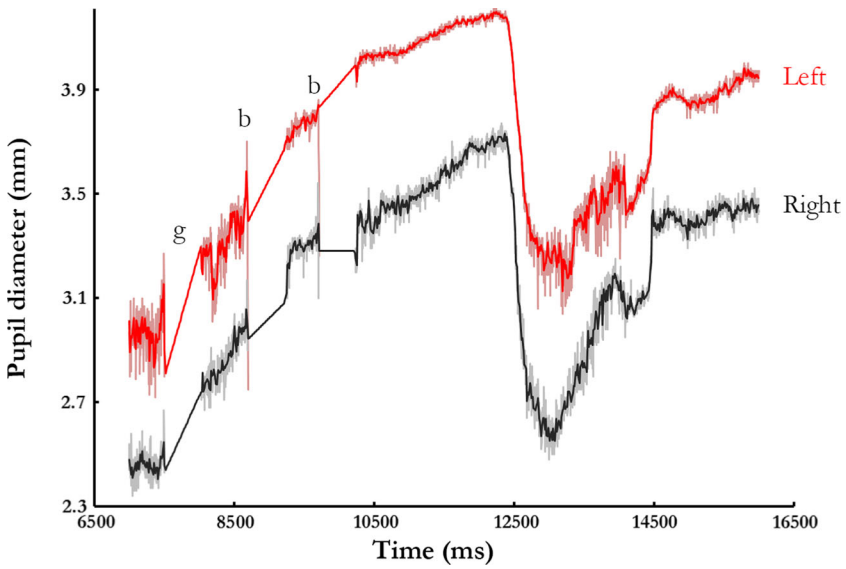

Fig. 3 Blink and gap correction in Pupilreduce 1.0. The total time interval is 1 s. The left and right pupil traces closely mimic each other, with a steady offset. The raw data are plotted in outer, lighter shading for each of the dark shaded lines. Averages of the four measurements are plotted in red (*left*) and black (*right*). Gaps (g) and blinks (b) were identified and discarded before linear interpolation

interpolated. Subsequently, the data were reduced by averaging four consecutive measurements. The eyetracker records the pupil diameters for both eyes. Measurements from each eye were treated separately, and the mean of the two values was used for the analysis. For each participant, pupillometry data were normalized according to the duration between the stimulus onset and response. This was done to ensure that the pupil diameter averages for multiple participants would average only equivalent processes (Piquado et al., 2010; Porter, Troscianko, & Gilchrist, 2007). If averaging is answer-locked, then someone who answers a question faster will have their initial pupil diameters averaged with the pupil diameters from the middle of the response of someone who takes longer. Similar problems would arise if the response is stimulus- (slide onset) locked.

The means of the pupil diameters from a 5-s exposure to the black and white slides were used to normalize the TEPR according to the following equation:

$$(\mathrm{PD} - \mathrm{D_{white}}) \Big/ (\mathrm{D_{black}} - \mathrm{D_{white}}) \times 100$$

where PD is the recorded pupil value for each of the question slides, D_{white} is the average value obtained from the white slide, and D_{black} is the average value obtained from the black slide. This is similar to the normalized protocol used by Piquado et al. (2010), in which they normalized according to the maximum and minimum pupil diameters. However, some of our normalized values were negative. This was most likely due to an underestimation of D_{black}, in our case, because of the brighter ambient-light conditions in our lab (250 lux) than in Piquado et al. (2010; Tepring Piquado, pers. comm.) The luminances of the white and black screens were 107 and 37 lux, respectively. For a truly dark screen, the luminance should be closer to zero. However, the net effect from such conditions is expected to be a small constant vertical shift in the normalized TEPR curve, and therefore, expected not to affect any of the conclusions of the study.

Results

Grades

All participants, except one, answered Q1. None answered Q1 incorrectly. Question 2 required substantial calculations, and the results were more varied. No one skipped Q2. Most of the answers included a response of 175 or a similar value at the beginning, but they differed widely in terms of their orders of magnitude, varying between 175 and 175,000. This indicated that participants understood the question and read the *y*-axis values correctly, but either did not know how to convert from bags to pounds or did not do that particular calculation correctly. Another common answer to Q2 was "January," indicating poor comprehension of the question. Four participants did not answer Q3: Two of them worked through the steps verbally and could not finish on time, and the other two simply did not answer the question. The answers to Q3 ranged from the low values of 225 to very high values of 325,000, with more intermediary numbers than were observed in the case of Q2, such as, 300, 2,600, 20,000, 22,500, 200,000, and so forth. This reflected the increased number of cognitive steps and their possible permutations, thereby increasing the number of unique values. The mean total score for participants' responses was 5.78 (*SD* = 2.21) (Fig. 4).

Higher accuracy of the answers was correlated with response time. The scores for Q2 and Q3 showed moderate correlations with response time (r = .48 and .39, respectively). The mean response times for Q1 to Q3 were 6.98 s (*SD* = 3.18), 21.29 s (*SD* = 7.72), and 26.83 s (*SD* = 7.26), respectively.

For further analysis, we separated high- and low-performing participants into two separate groups. Group 1 consisted of participants who scored equal to or greater than 7, and Group 2 consisted of those who scored less than or equal to 4. In this step, we excluded any participant who had not answered all of the questions. Therefore, Group 1 consisted of 11 participants, and Group 2 consisted of six (Fig. 4). The average response times (in seconds) for Q1 were comparable between Group 1 (*M* = 6.47, *SD* = 2.45) and Group 2 (*M* = 6.64, *SD* = 2.24). Group 2 (*M* = 16.66, *SD* =

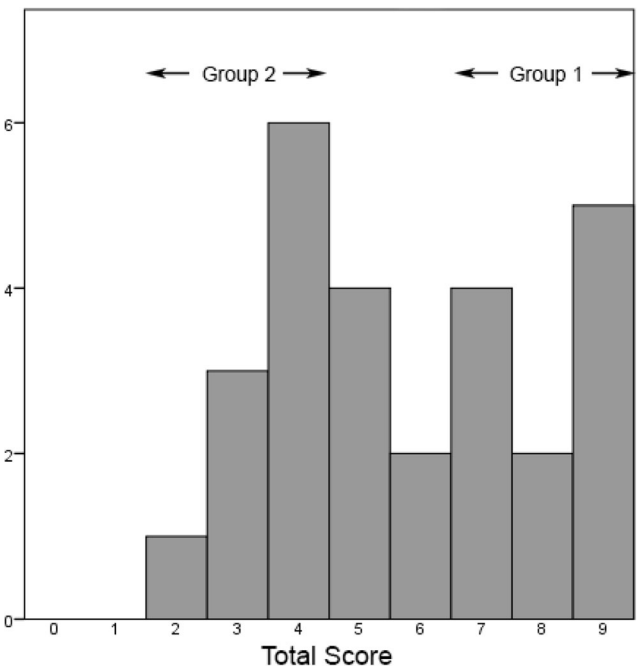

Fig. 4 Histogram of total scores for the two groups. All 11 candidates were retained for Group 1. Four of the ten potential candidates for Group 2 were excluded because they had not answered Q 3

8.09) answered Q2 faster than did Group 1 (M = 23.30, SD = 7.63). Similarly, Group 1 (M = 28.49, SD = 6.11) took more time to answer Q3 than Group 2 (M = 22.29, SD = 6.79).

Pupillometry

Figure 5 shows the mean pupil diameters for all participants, with standard error bars. In this analysis, all answers from all participants were included. Therefore, if a participant answered only Q1 correctly, answered Q2 incorrectly, and did not answer Q3, the data from Q1 and Q2 were included in the analysis. Individual pupil traces (Fig. 5, inset at top) showed an initial drop in pupil diameter to adjust to the white background of the question slides (which were preceded by gray slides). This is also evident in the average values, and more so for Q1 because of the shorter average response time; consequently, the averaging interval had less of a contribution from higher pupil diameters than those for Q2 and Q3. Therefore, excluding values at normalized time = 0, the TEPR demonstrates a steady increase for each question between 0.2 and 1 s, and subsequent declines. With increasing cognitive load, the pupil dilates until the moment the answer is verbalized. Subsequent contraction of the pupil indicates a dissipation of the cognitive load. Additionally, the standard errors of the means are markedly higher after normalized time point 1 (Fig. 5, inset at bottom), indicating that thought processes are less focused than the on-task thoughts prior to answering.

The mean TEPRs are also indicative of the cognitive demands of the respective tasks. Question 1 was the simplest, requiring only QC and IL. Even those two components would arguably be easier in Q1 than in Q2 and Q3. Question 2, on the other hand, required all three cognitive processes: QC, IL, and MC. Finally, Q3 was similar to Q2, but with an emphasized MC component. As expected, Q1 evoked the lowest mean TEPR. However, the mean TEPR for Q2 was consistently higher than that for Q3 in the second half of the task. This relationship is counterintuitive, because Q3 should demand substantially higher MC than Q2. It is important to note, however, that TEPR does not by itself signify the difficulty of a question, because decreased cognitive load for a longer duration can be an effective way to solve a difficult problem. The mean time of response for Q2 being less than Q3 also indicates that Q3 was likely more difficult to answer than Q2.

To further investigate this relationship, we performed a mixed-effects linear regression including a binary indicator of group status. This allowed us to test for the time

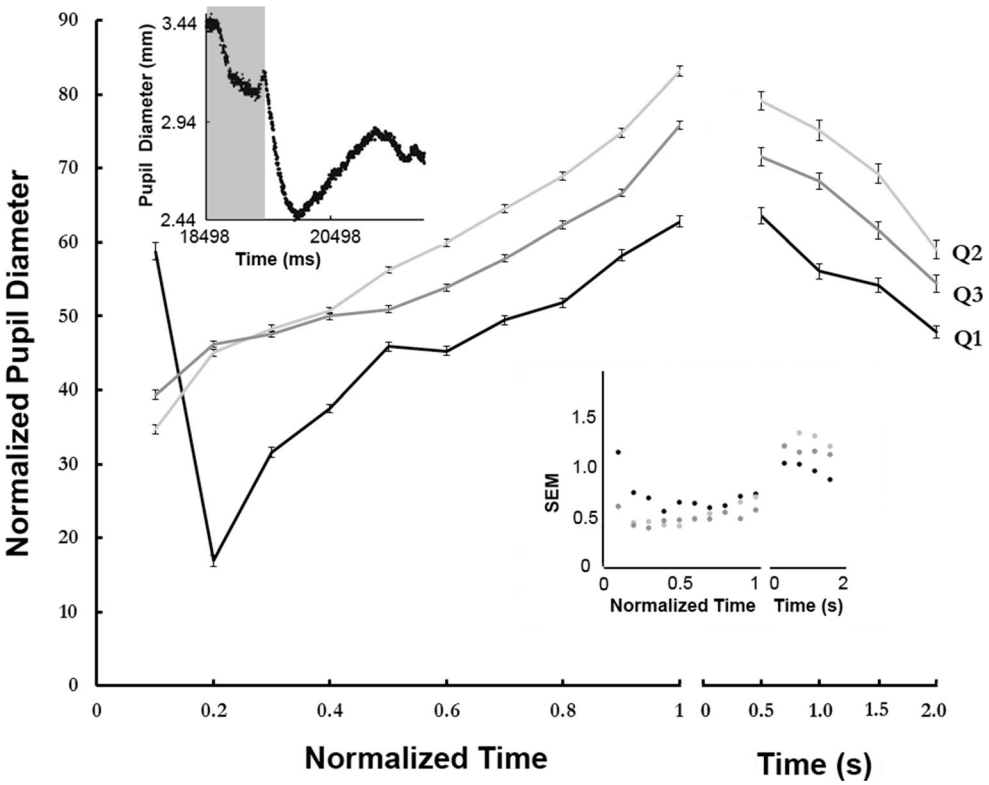

Fig. 5 Mean pupil diameter (normalized) of 27 participants. The *x*-axis is normalized by response time, and post-response time is expressed in seconds (break in the *x*-axis). The *top inset* shows the raw pupil diameter changes between a gray slide and a question slide (Q1). The shaded region indicates the duration of the gray slide. The sharp drop is a result of the change in brightness from the gray to the question slide, which is predominantly on a white background (see Fig. 1). This luminosity-controlled pupil diameter change causes the values at normalized time = .1 to be higher than would be expected from the task-evolved pupil diameter change alone. The *bottom inset* shows the standard errors of the means, plotted as a function of time

trends that might be different in each of the groups, by including interactions between group status and the time trend (Table 1, Fig. 6). The eyetracking data revealed that participants used a QC–IL–MC cognitive framework for the questions. In such a framework, we argue that the first half of the duration of a response is spent mostly in QC and IL, and the second half reflects more MC. Therefore, in our regression models we fit a piecewise linear regression with a break at .5 normalized time. If the slopes before and after the break did not differ significantly ($p > .05$), we used the simpler model—that is, without a break at .5. Incidentally, we also tried models with breaks at points other than .5, without any significant changes to the regression results. We included a participant-specific random intercept in the models to account for the repeated measures within each participant.

For Q1, we note a distinct break at normalized time = .5 in Group 2 that is absent in Group 1. In Group 2, Q1 shows a steeper initial slope than the other questions. The groups seem to respond to Q2 and Q3 differently. Group 2's TEPR confirms our initial expectation that TEPR is proportional to working memory requirements. Each successive TEPR, from Q1 to Q3, is higher. However, Group 1's TEPR is highest for Q2. Additionally, and more interestingly, Q3's TEPR has a flat initial slope for this group, the only segment across groups in which the TEPR does not increase with time.

Discussion

CLT is an instructional design paradigm that seeks to understand the cognitive-load requirements of instructional materials (Paas & Van Merriënboer, 1994a; Sweller, 1988; Sweller, van Merrienboer, & Paas, 1998; van Merriënboer & Sweller, 2005). On the basis of findings from memory research, CLT offers a framework to understand the role of working memory in problem solving, the relation of working memory to long-term memory, and the different types of cognitive loads (extrinsic, intrinsic, and germane) associated with instructional materials. Since TEPR reflects the cognitive loading of working memory, valuable insights can be gleaned by discussing the results within a CLT paradigm.

CLT is premised on the fact that working memory is limited in both capacity and duration. Working memory is believed to store approximately seven elements and can operate on two to four at one time, and most information is lost in less than 20 s (Miller, 1956; van Merriënboer & Sweller, 2005). However, working memory is easily bolstered by long-term memory, which has an unlimited capacity. During learning there is constant exchange of information between working memory and long-term memory. Relevant bits of information that can help with problem solving are grouped into schemata that are passed to the long-term memory by the working memory. Several

Table 1 Fitted regression parameters for each group, with individual task-evolved pupillometry response as the outcome variable

			Group 1 (Total ≥ 7)			Group 2 (Total ≤ 4)		
			Estimate	*SE*	Pr > |*t*|	Estimate	*SE*	Pr > |*t*|
Whole	Q1	Intercept	**−7.2**	**10.6**	**.5124**	−2.87	13.4	.8389
		Slope	**73.68**	**13.78**	**.0001**	68.85	17.1	.0002
	Q2	Intercept	**22.7**	**11.3**	**.0715**	**17.65**	**14.2**	**.2683**
		Slope	**50.19**	**12.4**	**.0001**	**42.22**	**16.9**	**.016**
	Q3	Intercept	41.7	11.9	.0057	**23.58**	**12.1**	**.1093**
		Slope	21.46	13.16	.1065	**47.71**	**15.2**	**.0029**
Piecewise	Q1	Intercept	−14.55	12.3	.2631	**−19.42**	**14.8**	**.246**
		Slope 0.2_0.5	101.41	27.2	.0003	**131.39**	**32.1**	**.0002**
		Slope_Diff	−44.12	37.3	.2398	**−99.28**	**43.8**	**.028**
	Q2	Intercept	23.46	12.2	.0826	13.03	15.7	.4447
		Slope 0.2_0.5	47.05	22.5	.0398	60.61	31.7	.0618
		Slope_Diff	5	29.9	.8677	−29.31	42.6	.4951
	Q3	Intercept	**52.26**	**12.7**	**.0021**	15.85	13.7	.2997
		Slope 0.2_0.5	**−22.33**	**23.3**	**.3407**	77.52	29	.0105
		Slope_Diff	**69.86**	**69.9**	**.0266**	−47.41	39.4	.2355

In the model, Slope 0.2_0.5 represents the slope of the line between normalized times .2 and .5—that is, prior to the breakpoint. Meanwhile, Slope_Diff represents the difference in the slope after the break at normalized time = .5 from the slope before the break. Hence, a nonsignificant result for Slope_Diff denotes that a single-line fit with no breakpoint is sufficient. Likewise, a significant result denotes that a breakpoint is a better fit. The solutions selected for analysis on the basis of this criterion are in bold

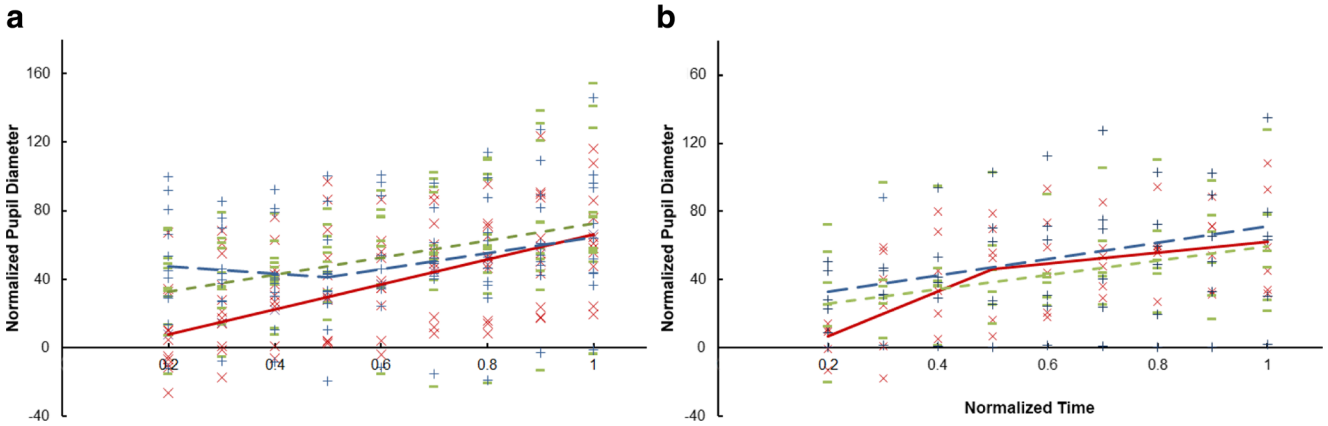

Fig. 6 Regression analysis of Groups 1 and 2. (a) Individual estimates from the 11 participants of Group 1 (Q1–*red crosses*, Q2–*green dashes*, Q3–*blue plus signs*) and the corresponding regression analysis (Q1–*red solid line*, Q2–*green line with small dashes*, Q3–*blue line with long dashes*). (b) Same for the six participants of Group 2

such schemata can be further grouped together, and crucial aspects of learning are building, storing, and automating schemata in long-term memory, so that when it is faced with a similar problem in the future, working memory can access the necessary schemata from the long-term memory. Schemata vary in their complexity, but they act like a single block of information, and therefore relieve working memory by reducing some of the cognitive load, since the operations within schemata do not require further working memory space.

An example will elucidate the concept of working memory, schemata, and long-term memory. When learning to drive for the first time, working memory is full of new sensory information, such as the next bend on the road, the traffic signal at an intersection, pedestrians, the car stereo volume, background chatter from co-passengers, and so forth. To reduce the working memory load, the driver might ask for the car stereo volume to be turned low and ask fellow passengers to remain silent. However, with more practice working memory seems to be more accommodative to these extraneous sensory inputs, which are not key to driving safely on the road. In terms of CLT, this phenomenon is simply proper management of the cognitive load experienced during driving, which comes with practice. All key elements relevant to safely driving the car gets arranged in schemata that are pushed into long-term memory. One such schema could be made up of the following distinct steps:

[
1) Look ahead, right, and left to note the traffic signal and crossing pedestrians.
2) If the light is green and no pedestrian is crossing, then drive ahead.
3) If Step 2 is false, then engage the brakes.
] SCHEMA

The schema above can be broken down into smaller subschemata. For example, engaging the brakes is itself a schema, involving disengaging the accelerator and pushing on the brakes (for an automatic car). With practice, relevant schemata get clustered together and stored in long-term memory, and can then be retrieved by working memory.

According to CLT, cognitive load can be categorized into three distinct types: extraneous, intrinsic, and germane (Sweller, 1988). *Extraneous* load refers to any cognitive load that arises from aspects of the problem presentation or design that are not critical to problem solving and are largely distracting. In the driving example, such loads would be those arising from the car stereo and chatter. *Intrinsic* load refers to that part of the cognitive load that is intrinsic to the problem. For example, driving on a crowded road will present a higher intrinsic load than driving on an empty street. *Germane* load refers to the part of the cognitive load that helps in schema formation and automation. Instructional design theory aims to reduce extraneous load and promote germane load. With more practice—that is, with more schema formation and automation—the intrinsic load, which is a function of problem difficulty and an individual's experience, reduces, and the working memory can then accommodate more extraneous load. This explains a driver's ability to focus on conversations or listen to music when driving on familiar roads.

The questions in this study were based on a simple figure illustrating the relationship between two variables. The visual content was assumed to provide a low and constant extraneous load for the two groups. The questions, however, were framed to have successively higher intrinsic load. Question 1 simply asked for the name of a month, Q2 required a single multiplication, and Q3

required three multiplications and addition of the products (or, alternatively, three additions and one multiplication). In the remaining discussion, we propose a CLT model (Fig. 7) that can qualitatively explain the results.

Group 2 indeed reflects the increasing intrinsic load of the questions, since TEPR is higher for each successive question. Furthermore, Q1 shows an initial steep slope (see Fig. 6). This is interpreted to be reflective of new information. As the graph was seen for the first time, the intrinsic load was most likely particularly high for Group 2. Even more suggestive is the fact that the slope shows a marked shallowing from normalized time = .5, indicating the realization of the simplicity of the problem. However, for the same question, Group 1 showed a gentle slope with no breaks, indicating lower overall intrinsic load, most likely arising from greater prior familiarity with such problems or higher working memory capacity. The distinct TEPR graphs of the two groups yet their very similar accuracies (all participants got Q1 correct) illustrate how learners can "compensate for an increase in mental load by investing more mental effort, thereby maintaining performance at a constant level" (Paas, Tuovinen, Tabbers, & Van Gerven, 2003, p. 67; see also Paas & Van Merriënboer, 1993). Since both the groups' performance remained indistinguishable and accurate, it is easy to mistake this accuracy for similar mental effort being exerted by the two groups. However, this result demonstrates that even for extremely simple questions such as Q1, the available information can often elicit different levels of mental effort. With increasing question difficulty, TEPR is likely to demonstrate further differences before the difference is evident from the accuracies of the learner responses. Therefore, pupillometry seems to be able to detect subtle differences in learner efficiency (or, conversely, instructional conditions, if learner efficiency remains constant), even before these are evident in student performance.

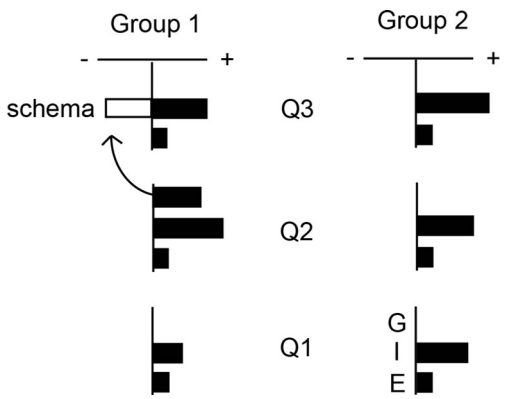

Fig. 7 A cognitive-load theory model. The contributions from extraneous (E), intrinsic (I), and germane (G) load are shown as *horizontal bars* for each question in the two groups. A solid bar indicates positive cognitive load and a blank bar indicates negative load. The total cognitive load of any question is given by the sum of the three bars. Extraneous load has been kept constant in this example. For Group 1 (left), intrinsic load is low initially and subsequently increases. Germane load appears only for Q2, which forms a schema that is passed to long-term memory. This schema reduces the cognitive load during Q3; the intrinsic load is less for Q3 than for Q2 because of question familiarity. In contrast, for Group 2 (right), each question poses successively higher intrinsic load. No schema is formed, and each question is treated in silo

Question 2 posed a higher intrinsic load because it required MC. Group 1 took longer to solve Q2 and also showed a steeper slope in TEPR, indicating substantially higher cognitive load. The markedly lower cognitive load observed in Group 2 could be due to poor engagement and/or an inability to identify the difference between Q1 and Q2. The latter processes would lead to lower mental effort because the question was incorrectly assumed to be simpler than it really was. Similarly, the higher cognitive load of Group 1 could arise out of greater motivation to be engaged with the material. Regardless of the cause behind the difference between Groups 1 and 2, both their performance and the TEPR results suggest deeper cognition in Group 1. The substantially higher cognitive load (between Q1 and Q2 within Group 1) is what would be expected from van Merriënboer and Sweller (2005), where the authors discuss how variability in question difficulty promotes schema formation (p. 161). The relative difficulty of Q2 juxtaposed to the ease of Q1 is ideal for such schema formation. Therefore, part of the cognitive load observed for the Q2 TEPR was germane load that helped in schema formation (see Fig. 6).

The TEPR results for Q3 adds further credence to this argument. Group 1 shows a flat TEPR initially, as well as a TEPR that is lower than for Q2. Together, these facts suggest that some schema formation must have taken place during Q2 that was then used to understand and solve Q3. The steps that are required to solve Q3 are only an extension of those in Q2. Instead of a single multiplication for a single month, Q3 requires three such multiplications and the addition of the products. For example, a successful schema formed during Q2 could be the following:

[
1) Look up number of bags for month
2) Multiply by the conversion factor
] SCHEMA

A solution for Q3 would then require the operation SCHEMA + SCHEMA + SCHEMA.

Therefore, the initial reduction of cognitive load can be ascribed to the successful categorization of Q3 as an extension of Q2, thereby leading the participants to expect lower cognitive load. Changes in pupil diameter in anticipation of the task to be performed have previously been noted by other authors (Kahneman & Beatty, 1966; Piquado, Isaacowitz, & Wingfield, 2010). Engagement of

the schema to actually solve the problem would only occur during the second half of the problem-solving process, during which the TEPR slope was less than that for Q2. However, it must be remembered that the average time of response was greater for Q3 than for Q2. Therefore, we interpret the combined results of response time and the TEPR slopes of Q3 and Q2 thus: The TEPR of Q3 was more likely indicative of a known but difficult task, as compared to Q2, which was unknown and required more cognitive load.

In contrast, Group 2 formed no successful schema during Q2. Thus, the extra load observed for Q3 resulted from both a miscategorization of the problem (Q3 being thought of as distinct from Q2) and/or the lack of a successful schema in addressing the problem.

The results show that in addition to the original three criteria for a suitable proxy for cognitive load (Beatty, 1982; Kahneman, 1973), pupillometry has the potential to be used as a tool to measure learning. In this context, if more-difficult tasks evoke lower pupil diameters than do less-difficult tasks, then regardless of the total time taken to complete the tasks (which is still expected to reflect the computational difficulty), some form of learning must have taken place. In terms of working memory considerations, this decrease in pupil diameter probably reflects the part of working memory that remains tied up to figure out the right way to approach the task and/or the part of working memory that is not free due to the inherent uncertainty from the open-endedness of the task. Only after successful completion of a task and recognition that a subsequent task is similar to a prior task is the working memory able to "free up" this part. Since, successful problem categorization is an important component of expertise or learning (Chase & Simon, 1973; Chi, Feltovich, & Glaser, 1981; de Groot, 1965), pupillometry can be an effective way to study either of these constructs.

For instructional design, the results provide an important insight. Supporting earlier studies, schema formation was found to be an important step toward problem solving (Gick & Holyoak, 1983). It seems that some students are better at schema formation than others. We also observed that the group that successfully formed schemata did so spontaneously and showed a peak in cognitive load when two questions differed highly in their difficulty levels (Q1 and Q2). The same group showed a drop in cognitive load when two similar problems were juxtaposed (Q2 and Q3). Therefore, at least in some participants (Group 1), the ability to successfully solve problems arises partly from their ability to identify differences and similarities between the problems, a classic expert trait. In others (Group 2), this ability may not be spontaneous for a variety of reasons, including low cognitive ability, disengagement, and so forth. Regardless, it might be good instructional-design practice to explicitly juxtapose high-variability concepts, as Paas and Van Merriënboer (1994b) and van Merriënboer, de Croock, and Jelsma (1997) have suggested. Additionally, more research needs to be conducted to understand whether explicit statements about problem differences and similarities that accompanied a connected set of instructional elements could raise the performance of Group-2-like participants.

Limitations

Given that our total number of participants in the study was 29, and that we had further examined 17 of these 29 in our subgroup analysis, some of our results that were not statistically significant could have been due to lack of power. Thus, although we did not find significant break points in some instances, further investigation is warranted.

Care was taken to ensure minimal preloading of participant expectations. A pilot study showed that students were not reading Q2 and Q3 carefully, and thus were assuming them to be similar to Q1 and answering accordingly. Since Q1 did not require any MC, we found that the performance of students improved after they were told to expect some MC after Q1. However, they remained unaware of the relative difficulty of the two remaining questions. Furthermore, they were told that the time allocated for each question would be sufficient for them to be able to answer it—that is, less time for easier questions and more time for difficult questions. This direction was given to reduce the possibility of any inadvertent performance stress. However, despite our taking the precautions above, some preloading of expectation and/or question dependent stress cannot be completely ruled out.

To ensure that the pupil data could be analyzed meaningfully, we needed to minimize look-away time from the monitor screen. As typing on a keyboard or writing with a pencil involves such gaze stray (to the keyboard or the piece of paper), we asked the participant to verbalize the answers while staying focused on the screen. Verbalization of intermediate steps in the calculations was strictly discouraged, because that might lead to a drop in pupil diameter (Piquado et al., 2010; Porter, Troscianko, & Gilchrist, 2007) through a reduction of working memory load. These experimental conditions might not have been representative of authentic learning experiences.

We assessed cognitive load with compound, interdependent tasks based on a three-item problem set. Although, the cognitive processes in the three questions were similar and can be broken down into the QC–IL–MC scheme, with Q1

requiring QC–IL and Q2 and 3 requiring QC–IL–MC, the three items might not have sufficiently captured the full gamut of cognitive processes involved in such problem solving. Moreover, the generalizability of our findings can only be tested with a validated and reliable instrument that includes similar problems, preferably with and without graphical inputs.

Pupil foreshortening error (PFE), or the error in pupil diameter measurement that covaries with gaze location, was a possible confound in this study (Brisson et al., 2013; Gagl, Hawelka, & Hutzler, 2011). Such errors are most problematic when they are systematic. For example, in a reading task the stimulus (the paragraph) runs from top left to bottom right, which might create a problem, because the pupil diameters could be systematically overestimated at the top left corner and underestimated at the bottom right, or vice versa (Brisson et al., 2013), thereby leading to a pupillometry profile that would show a gradual decrease (or increase) regardless of the cognitive-load profile of the passage. Such errors can lead to Type I errors (Hayes & Petrov, 2016). In this study, the areas of interest for each task were spread across the screen, and the visits/revisits to these areas by the participants were haphazard. Furthermore, between group comparison of gaze positions revealed no significant difference. Therefore, systematic errors arising from PFE is expected to be negligible in this study. However, some amount of PFE cannot be ruled out and likely added to the noise in the dataset.

Conclusion

The overarching goal of this study was to demonstrate the utility of the combined use of pupillometry and eyetracking data, which fall under distinct epistemological frameworks. Research questions in education are multifaceted and comprise tasks that have several components. For example, in this study we looked at graphical problem solving, which comprises problem comprehension, information look-up, and problem solving. To understand cognitive load during graphical problem solving, we need to identify the cognitive load from each of the components or subtasks, and their subtle interactions. Traditional pupillometry methodology offered no scope for such observations, because pupillometry data cannot constrain subtask distribution of the cognitive load. The combination of gaze data from eyetracking with traditional pupillometry data can help us overcome this limitation. We demonstrated that under certain assumptions and a careful experimental design, such a methodology can provide new insights in education research.

The methodology in this study can be extended to several other contexts. For example, consider a map-reading task. The viewer must understand the legend and scale, and must use that information to locate places or estimate distances between two points on a map. If we intend to understand the cognitive load from reading a map, then that task itself would be composed of several subtasks, such as legend and scale look-up, search for locations, and mental calculation. The use of gaze data to identify these individual subtasks and the analysis of pupillometry data within this framework can provide answers to interesting research questions, including a better understanding of which aspects of the map produce greater-than-expected cognitive load and may need modification (see also Klingner, 2010).

The following are important conclusions from this study:

1. We demonstrated for the first time a pupillometry profile in graphical problem solving. Similar to earlier studies with simpler tasks, like counting and searching, the within-task profiles reflected a progressive increase of cognitive load with time. The peak mental load was achieved before the answer was verbalized. Subsequently there was a drop in pupil diameter, indicating the freeing up of working memory.
2. Unlike previous studies in which the tasks were independent, this study revealed that pupillometry is also sensitive to the interdependency of tasks. This was indicated by a lower TEPR for Q3 than for Q2 in some of the participants (Group 1), when in fact Q3 required more mental calculation. However, the biggest difference between this study and earlier studies was the compound nature and interdependence of tasks. The observed TEPR suggests that pupillometry can signify mental effort in an even more holistic manner than has previously been demonstrated. Regardless of the effort put toward MC, for which total duration and percent correct would be equally good proxies, pupillometry can the indicate ability to categorize problems, a key aspect of learning.
3. In terms of CLT, this study provides an advance over earlier studies. CLT depends critically on the tripartite subdivision of total cognitive load into intrinsic, extrinsic, and germane loads. However, measurement of these individual constructs has been fraught with difficulty (Paas et al., 2003, p. 67). There are contradictory evidence on the independency of the subdivisions and their measurement (see, e.g., Kalyuga, 2011, and DeLeeuw & Mayer, 2008). The results presented in this study support a tripartite CLT model, in general, and the existence of germane load as an independent construct, in particular.

Acknowledgements This material is based on work supported, in part, by the National Science Foundation under Grant No. DRL-1443024. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We thank Tepring Piquado for her comments, and also thank the participants who took part in this study.

References

Ahern, S., & Beatty, J. (1979). Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science, 205,* 1289–1292.

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

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New YorkY: Academic Press. doi:10.1016/S0079-7421(08)60452-1

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91,* 276–292. doi:10.1037/0033-2909.91.2.276

Beatty, J., & Kahneman, D. (1966). Pupillary changes in two memory tasks. *Psychonomic Science, 5,* 371–372.

Ben-Nun, Y. (1986). The use of pupillometry in the study of on-line verbal processing: Evidence for depths of processing. *Brain and Language, 28,* 1–11.

Brisson, J., Mainville, M., Mailloux, D., Beaulieu, C., Serres, J., & Sirois, S. (2013). Pupil diameter measurement errors as a function of gaze direction in corneal reflection eyetrackers. *Behavior Research Methods, 45,* 1322–1331.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55–81.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121–152.

Cierniak, G., Scheiter, K., & Gerjets, P. (2009). Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Computers in Human Behavior, 25,* 315–324.

de Groot, A. D. (1965). *Thought and choice in chess*. The Hague: Mouton.

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology, 100,* 223–234.

Gagl, B., Hawelka, S., & Hutzler, F. (2011). Systematic influence of gaze position on pupil size measurement: Analysis and correction. *Behavior Research Methods, 43,* 1171–1181. doi:10.3758/s13428-011-0109-5

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15,* 1–38. doi:10.1016/0010-0285(83)90002-6

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology, 33,* 457–461.

Hayes, T. R., & Petrov, A. A. (2016). Mapping and correcting the influence of gaze position on pupil size measurements. *Behavior Research Methods, 48,* 510–527. doi:10.3758/s13428-015-0588-x

Heinrich, W. (1896). Die Aufmerksamkeit und die Funktion der Sinnesorgane: Erster Beitrag. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane, 9,* 342–388.

Hess, E. H. (1972). Pupillometrics: A method of studying mental, emotional and sensory processes. In N. S. Greenfield & R. A. Sternbach (Eds.), *Handbook of psychophysiology* (pp. 491–531). New York: Holt, Rinehart & Winston.

Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science, 132,* 349–350.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science, 143,* 1190–1192.

Hyönä, J., Tommola, J., & Alaja, A. M. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *Quarterly Journal of Experimental Psychology, 48,* 598–612.

Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology, 47,* 310–339. doi:10.1037/h0078820

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs: Prentice Hall.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science, 154,* 1583–1585.

Kalyuga, S. (2011). Informing: A cognitive load perspective. *Informing Science, 14,* 33–45.

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology, 93,* 579.

Klingner, J. (2010). Fixation-aligned pupillary response averaging. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (pp. 275–282). New York: ACM.

Komogortsev, O. V., Gobert, D. V., Jayarathna, S., Koh, D. H., & Gowda, S. M. (2010). Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on Biomedical Engineering, 57,* 2635–2645. doi:10.1109/TBME.2010.2057429

Körner, C. (2011). Eye movements reveal distinct search and reasoning processes in comprehension of complex graphs. *Applied Cognitive Psychology, 25,* 893–905.

Körner, C., Höfler, M., Tröbinger, B., & Gilchrist, I. D. (2014). Eye movements indicate the temporal organisation of information processing in graph comprehension. *Applied Cognitive Psychology, 28,* 360–373.

Landgraf, S., Van der Meer, E., & Krueger, F. (2010). Cognitive resource allocation for neural activity underlying mathematical cognition: A multi-method study. *ZDM, 42,* 579–590.

Leppink, J., Paas, F., Van der Vleuten, C. P., Van Gog, T., & Van Merriënboer, J. J. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods, 45,* 1058–1072. doi:10.3758/s13428-013-0334-1

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97. doi:10.1037/h0043158

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84,* 429–434. doi:10.1037/0022-0663.84.4.429

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. G. W. C., & 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. doi:10.1177/001872089303500412

Paas, F. G. W. C., & Van Merriënboer, J. J. G. (1994a). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review, 6,* 351–371.

Paas, F. G. W. C., & Van Merriënboer, J. J. G. (1994b). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86,* 122–133. doi:10.1037/0022-0663.86.1.122

Peavler, W. S. (1974). Pupil size, information overload, and performance differences. *Psychophysiology, 11,* 559–566.

Pinker, S. (1990). A theory of graph comprehension. In R. O. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp. 73–126). Hillsdale: Erlbaum.

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology, 47,* 560–569.

Porter, G., Troscianko, T., & Gilchrist, I. D. (2007). Effort during visual search and counting: Insights from pupillometry. *Quarterly Journal of Experimental Psychology, 60,* 211–229.

Schiff, J. M. (1875). *La pupille considérée comme esthésiomètre* (J. B. Baillière, Ed.). Paris: J.-B. Baillière et fils.

Schmidtke, J. (2014). Second language experience modulates word retrieval effort in bilinguals: Evidence from pupillometry. *Frontiers in Psychology, 5,* 137. doi:10.3389/fpsyg.2014.00137

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review, 14,* 47–69.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12,* 257–285.

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10,* 251–296. doi:10.1023/A:1022193728205

Van Gerven, P. W., Paas, F., Van Merriënboer, J. J., & Schmidt, H. G. (2004). Memory load and the cognitive pupillary response in aging. *Psychophysiology, 41,* 167–174.

van Merriënboer, J. J. G., de Croock, M. B. M., & Jelsma, O. (1997). The transfer paradox: Effects of contextual interference on retention and transfer performance of a complex cognitive skill. *Perceptual and Motor Skills, 84,* 784–786.

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. doi:10.1007/s10648-005-3951-0

Whelan, R. R. (2007). Neuroimaging of cognitive load in instructional multimedia. *Educational Research Review, 2,* 1–12.