Assessing Cognitive Load in Adaptive Hypermedia Systems: Physiological and Behavioral Methods*

Holger Schultheis[1] and Anthony Jameson[2]

[1] Dept. of Computer Science, Saarland University, 66123 Saarbrücken, Germany, schulth@studcs.uni-sb.de
[2] DFKI, Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany, jameson@dfki.de

Abstract. It could be advantageous in many situations for an adaptive hypermedia system to have information about the cognitive load that the user is currently experiencing. A literature review of the methods proposed to assess cognitive load reveals: (1) that pupil size seems to be one of the most promising indicators of cognitive load in applied contexts and (2) that its suitability for use as an on-line index in everyday situations has not yet been tested adequately. Therefore, the aim of the present study was to evaluate the usefulness of the pupil size index in such situations. To this end, pupil diameter and event-related brain potentials were measured while subjects read texts of different levels of difficulty. As had been hypothesized, more difficult texts led to lower reading speed, higher subjective load ratings, and a reduced P300 amplitude. But text difficulty, surprisingly, had no effect on pupil size. These results indicate that pupil size may not be suitable as an index of cognitive load for adaptive hypermedia systems. Instead, behavioral indicators such as reading speed may be more suitable.

1 Introduction

1.1 Assessing Cognitive Load for Adaptive Hypermedia Systems

There are many situations in which it would be useful for an adaptive hypermedia system to be able to assess the current cognitive load of the user. For example, suppose that the system notices that the current user is experiencing high cognitive load while reading a particular page. The system might then (a) insert more explanations and examples, (b) select as subsequent pages some pages that are inherently easier to read, or (c) eliminate unnecessary distractions (e.g., background music). Similarly, if the user's current cognitive load is lower than an optimal level, the system might increase its density of information presentation.

In some cases, prediction of cognitive load may be possible on basis of the page's intrinsic difficulty and the user's level of knowledgeability with respect to the subject matter; this type of estimation is commonly made in intelligent tutoring systems. But since such predictions cannot be entirely precise and reliable, it might be useful to have a more direct way of assessing cognitive load. In particular, it is desirable to obtain

* We thank Boris Brandherm for many fruitful discussions and the research group of Axel Mecklinger for invaluable support in the EEG part of the study.

load estimates which are fairly time-specific, so that the system can adapt quickly to a change in the users state.

But assessing the cognitive load of a given user is, unfortunately, in itself a difficult task. As a result, a range of different load assessment techniques has been proposed (see e.g., [1, 2]) over the years. The overall aim of the work presented here was to find the assessment method that seems best suited for building a user-adaptive system that utilizes information about the cognitive load of the user on-line while the user is reading text presented on a computer screen. In the rest of this section, the most important classes of assessment methods are discussed briefly with regard to their appropriateness as on-line measures of cognitive load. The remainder of the paper reports on and discusses an experiment in which an especially promising technique—measurement of pupil diameter—was evaluated.

1.2 Measures of Cognitive Load

Although there exist many methods for the assessment of cognitive load, each method can be assigned to one of four classes: 1. analytic measures, 2. subjective measures, 3. performance measures, and 4. psychophysiological measures.

1. As has already been mentioned, load estimation can be based on general (i.e. not interaction-specific) information about the system and the user(s). For example, information about the intrinsic difficulty of a hypertext page and about the expertise of the user working on this page can form a basis for the prediction of the user's load. But as this technique, which relies heavily on prior knowledge, does not take into account information about the current interaction, unforeseeable situations and individual peculiarities may lead to suboptimal system behavior.

2. Subjective measures involve questions which ask the user to rate the cognitive load that she has experienced or is experiencing. For example, a scale for a self-report of cognitive load could appear on each page of the hypertext system. But subjective reports are distorted by memory and consciousness effects. Moreover, administering the scale(s) after a certain page has been read would not enable the system to react to the user's needs while she is still reading that page. Asking the user for a report during the reading of a page, on the other hand, could be distracting.

3. With the third group of methods, the user's cognitive load is inferred from her overt behavior, or *performance*. A piece of evidence of this type might be, for instance, the speed with which the user reads the hypertext: This technique may not reflect all of the variations in cognitive load, and it is appropriate only if the activity yields a sufficiently high rate of observable behavior. These disadvantages can be avoided through the introduction of a *secondary task*. For example, the reader of the hypertext might be given the task of attending to a flashing light and pushing a button when a certain pattern of flashes occurs. Poor performance on the secondary task, for example, indicates that the primary task of reading induces high cognitive load. Although this approach in part avoids the disadvantages just mentioned, introducing a secondary task may itself be problematic, because it may disturb the user's main activity.

4. Finally, changes in various bodily processes and states are observed to covary with changes in cognitive load. Therefore, monitoring of these body functions sometimes allows load to be inferred. A major advantage of psychophysiological measures is

the continuous availability of bodily data, which potentially allows load to be measured with a high rate and high degree of sensitivity. What is more, without the introduction of an extra task, information about cognitive load is available even in situations in which overt behavior is relatively rare. Consequently, compared with the other classes, psychophysiological methods seem to be especially promising for on-line assessment in adaptive hypermedia systems.

Unfortunately, with many of the existing psychophysiological measures on-line assessment in an applied context is not currently feasible. Most of them require electrodes to be attached to the body (e.g., electroencephalogram, electrocardiogram, muscle tension) or the use of equipment that entirely rules out deployment in everyday situations (e.g., functional Magnetic Resonance Imaging, Positron Emission Tomography, magnetoencephalogram). Others, again, seem to be too indirectly linked to cognitive load (e.g., blink rate, blink duration) or to be too slow for on-line measurement (e.g., hormone level). In contrast, the measurement of the varying size of a person's pupil has none of these disadvantages. Not only is it regarded as "one of the most sensitive workload measures available" [3], but it is also assumed to respond to changes in load within several hundred milliseconds [4, 5]. Moreover, it is not necessary to attach any electrodes or other equipment to the user: The measurement can be accomplished with a remote eye tracker, which can be placed near the computer monitor. In sum, in addition to the advantages typical of all psychophysiological indicators, the measurement of pupil size has a combination of properties that seem uniquely well suited for the assessment of cognitive load during the use of adaptive hypermedia systems.

Over the last 40 years, a lot of studies have demonstrated the sensitivity of a person's pupil size to their cognitive load in a wide variety of tasks (see, e.g., [4] for a review): The higher the load, the bigger the pupil. In general, in these experiments at least two tasks of different difficulty were employed; subjects had to perform the tasks while their pupil diameter was recorded. The actual tasks used in these studies include, to name only a few: memorizing 3 vs. 7 digits [4]; shadowing words vs. translating them [6]; reading syntactically simple vs. complex sentences [7]; and telling the truth vs. lying [8]. These studies consistently reported larger pupil diameters during more difficult tasks.

But since pupil size is also especially sensitive to a number of influences not related to cognitive load (e.g., ambient light), previous works utilizing pupil size as a cognitive load indicator all used at least three of the following five means to control those influences: (a) constant lighting; (b) avoidance of eye movements; (c) use of nonvisual (e.g., acoustic) stimuli; (d) use of many similar, short tasks; and (e) evaluating only mean values averaged across tasks and subjects.

Such strict control of the environment is not realistic in connection with an adaptive hypermedia system; and averaging over tasks and subjects is not suitable for diagnosing the current load of a single person. Thus, to be truly useful in the situations of interest to us, pupil size should be a good indicator even if some or all of the above constraints are relaxed. To find out whether this result can be obtained, we designed and conducted a new experiment, which is discussed in the rest of this paper.

1.3 Measures Used in the Experiment

The aim of the experiment was to evaluate the utility of pupil size as an on-line measure of cognitive load for an adaptive hypermedia system, not actually to employ it in such an environment. As a result, we drew on additional techniques—some of which are not appropriate for applied contexts (see 1.2)— to provide information about the load. More precisely, behavioral, subjective and ERP measures were used. Whereas the first two methods are fairly straightforward, the third one may require some explanation, which will be given in the following paragraphs.

Processing of stimuli is accompanied by changes in brain activity, that is, activation or inhibition of certain neuronal ensembles. This neural activity is mainly electric, and it therefore generates electrical fields. These fields extend to regions outside the skull and can be recorded around the head. In particular, certain environmental events (e.g., a sound or a flash of light) give rise to characteristic and consistent variations in the electrical field around the head. These variations recorded from the scalp via electrodes are termed *event-related brain potentials* (*ERPs*). With regard to cognitive load measurement, one particular ERP, the P300, is of special interest. In general, this potential is elicited when a low-probability task-relevant stimulus is encountered (i.e., a stimulus to which the subject is attending). Moreover, it has been shown (see, e.g., [9]) that the more attention (mental effort) is devoted to the task associated with the evoking stimulus, the higher is the P300 amplitude.

The most common procedure up to now for utilizing this property of the P300 in load assessment has been to introduce a secondary task containing stimuli that elicit the P300. The magnitude of the evoked P300 gives information about the cognitive load in the main task: The larger the amplitude, the smaller the load. But with this method the subject is required to perform a secondary task, which may cause the same problems as those associated with the secondary task measure (1.2). To circumvent these disadvantages, we applied a different, relatively new technique that relies on the *Novelty-P300* [10]. This special subtype of the P300 is elicited by highly unexpected, previously unexperienced (i.e., novel) stimuli even if these stimuli are not attended to. As a result, the evoking stimuli do not have to be embedded in a secondary task. Instead, the Novelty-P300 can be elicited by a sequence of stimuli which are (a) presented simultaneously with the task of the user but (b) not relevant to that task. Regarding cognitive load, the Novelty-P300 has the same properties as the original P300: As [11] have shown, the Novelty-P300 is smaller for higher load. As in the approach described in [11], in our experiment P300s were elicited by sequences of sounds (cf. Sect. 2).

2 Method

Material. As material to be read at the computer by each subject, we prepared 8 texts—4 easy and 4 difficult—of approximately equal lengths. (Easy and difficult texts averaged 342 and 339 words, respectively, in length.) Difficulty was determined through subjective assessment and confirmed objectively in terms of the sources of the texts: Easy texts were taken from schoolbooks for the fifth grade and from children's books, while difficult texts were taken from schoolbooks for the 12th grade and from philosophical treatises. Text sequence was pseudopermuted via the Latin squares approach.

Given the fixed scheme ABBABAAB, where A and B denote difficulty classes, and a fixed order of texts within each class, different sequences were constructed via rotation of the texts of each class through the indicated positions.

Participants. Thirteen subjects, 8 female and 5 male, took part in the experiment. Their ages ranged from 20 to 41 years, with a mean age of 25.5 years. All were native speakers of German. They received either course credit or a monetary reward for their participation. In particular, to motivate careful reading, we paid subjects an extra reward of € 0.20 for each content question (see below) that they answered correctly.

Procedure. Subjects were seated facing a computer screen located at a distance of approximately 50 cm. For control of illumination, no external light was allowed to enter the room. The task of the participants was to read on the computer screen the texts described above. Presentation of each text comprised five phases: First, to produce a baseline value for pupil diameter, subjects were asked to fixate for 20 s a circle in the middle of a screen of Xs that had been arranged like the letters in normal text. Then, a real text was shown. Participants read the text at their own pace until they felt that they had understood it.[3] Then, four 7-alternative multiple-choice questions about the content of the text were to be answered. Finally, subjective ratings of text difficulty and the subject's own willingness to be interrupted were elicited as subjective load indicators.

Pupil size and point of gaze were measured throughout the whole experiment. In contrast, ERPs could be recorded only in the presence of eliciting tones, which were presented only during the actual reading of the texts. Besides, reading speed and number of correct responses were computed as behavioral measures of cognitive load.

Technology. Pupil diameter and point of gaze were recorded at 50 Hz with an ASL 504 remote eye tracking system that used pan/tilt optics.

In addition to vertical and horizontal electrooculograms, EEG was registered from 62 electrodes at a sampling rate of 500 Hz.[4]

For checking the luminosity of each text as displayed on the computer screen, a Gossen Lunasix F light meter was employed. Measurements indicated that luminosity was equal for all texts and the baseline screen.

ERPs were elicited by different types of tones played in random sequences to the subjects through speakers positioned to the left and right of the computer screen. Every 550 ms, a *standard*, a *deviant*, or a *novel* tone was presented for 200 ms with probabilities of 0.8, 0.1, and 0.1, respectively. Standard tones were 600 Hz sinus tones, deviant tones were 660 Hz sinus tones and novel tones were unique, nonsinus sounds (e.g., a honking sound) that were expected to evoke the Novelty-P300. With respect to the five means of control mentioned in 1.2, our setup led to the following relaxations: (a) use of visual stimuli (the texts); (b) occurrence of eye movements; and (c) use of relatively few but long tasks.

[3] In fact, reading time for each text was limited to 5 minutes, but no subject exceeded this time.

[4] For those interested in the details of this method: The electrodes were arranged according to the 10–10 system. Measurements took place referenced to the left mastoid with the forehead serving as ground and electrode impedance below $10\,k\Omega$. Signals were filtered on-line with a 0–70 Hz bandpass and a 50 Hz notch.

3 Results

Except where otherwise stated, the analyses reported in the following paragraphs are repeated-measures analyses of variance. Where appropriate, statistical significance was determined after correction of the degrees of freedom using Huynh-Feldt epsilon. The level of significance for all reported analyses was set to $\alpha = 0.05$.

Behavioral Data. More difficult reading, as [7] have shown, leads not only to higher load as indexed by the pupil but also to a smaller number of correct responses and slower reading. Accordingly, a lower reading speed and a lower number of correct responses for difficult texts were hypothesized for the current study. With respect to reading speed, this hypothesis was confirmed statistically ($F_{(5,62)} = 30.08,\ p < 0.001$, see Fig. 1a). This was not the case for the number of correct responses. Although in our data the answers to questions about difficult texts were less often correct than answers referring to easy texts (see Fig. 1b), a statistical comparison using the McNemar test revealed no significant difference ($\chi^2 = 2.64,\ p > 0.1$).

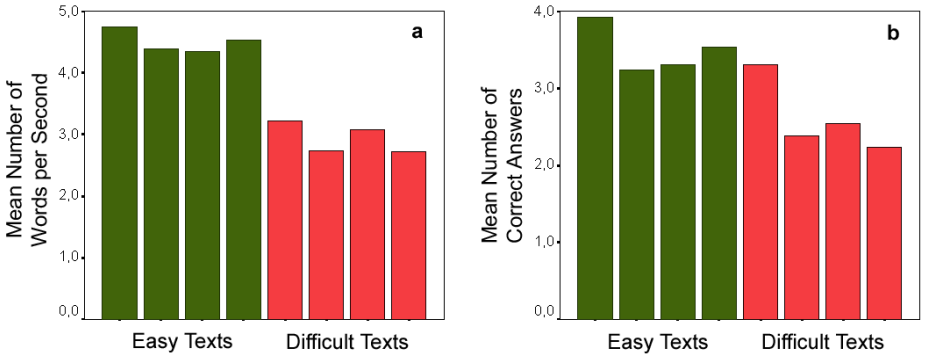

Fig. 1. (a) Mean reading speed in words per second and (b) mean number of correct answers for each of the eight texts.

Subjective Data. The subjective ratings of load consisted of judging on a 4-point scale both the experienced difficulty (1 = "easy" – 4 = "difficult") and how annoying an interruption during reading would have been (1 = "no problem" – 4 = "very annoying"). As expected, difficult texts were judged to be significantly more difficult ($F_{(7,84)} = 42.58,\ p < 0.001$, see Fig. 2a) and lower in terms of interruptibility ($F_{(7,84)} = 27.97,\ p < 0.001$, see Fig. 2b) than easy texts.

ERP Data. Since the Novelty-P300 is assumed to be especially pronounced over the upper forehead and the center of the scalp (see, e.g., [10]), examination was confined to two electrodes at these locations.[5] The first step of the analysis was to visually study the

[5] To be precise, these electrodes were FCz and Cz according to the 10–10-system.

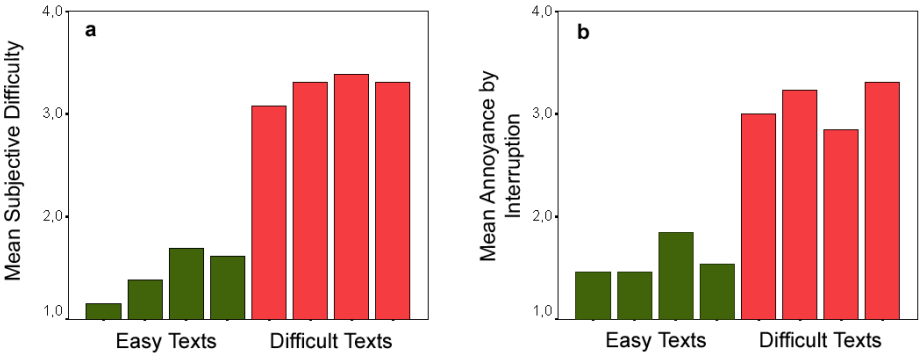

Fig. 2. Part (a): Mean subjective difficulty ratings, ranging from ”easy” (= 1) to ”difficult” (= 4) for all eight texts. Part (b): Mean annoyance-by-interruption ratings ranging from ”no problem” (= 1) to ”very annoying” (= 4) for each of the eight texts.

electrooculogram recordings so as to reject or correct trials that showed eye movement artifacts or blink artifacts. From the resulting trials, for each subject four average curves (curves evoked by standard and novel sounds while reading easy or difficult texts) were built, which, collapsed over participants, resulted in the grand average waves displayed in Fig. 3. P300 amplitude was then defined as the local maximum of the difference curve—obtained by subtracting easy/difficult standard curves from the corresponding novel curves—in the time from 164 to 274 ms after stimulus onset. In accordance with theory, the P300 amplitude at the two electrodes was significantly larger (one-tailed) during the reading of easy as opposed to difficult texts ($F_{(1,12)} = 3.5,\ p < 0.05$). In other words, the ERP method revealed a higher cognitive load while reading difficult vs. easy texts.

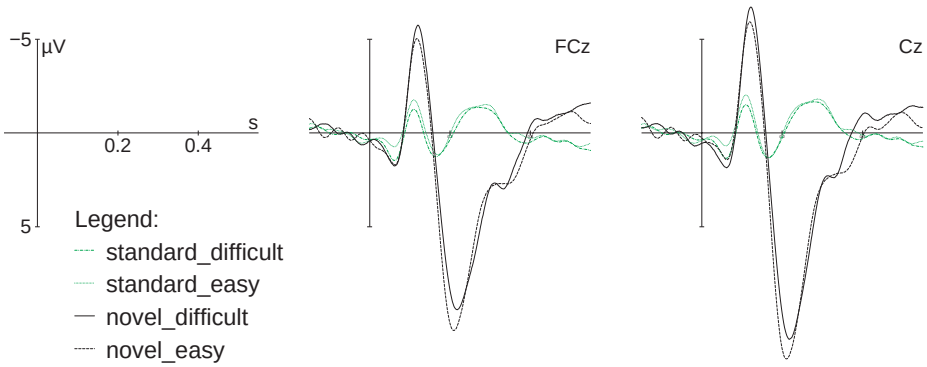

Fig. 3. Grand average ERPs elicited by standard and novel sounds while reading easy and difficult texts.

Pupil Data. As with the ERP data, prior to further analyses eye blinks had to be identified in and eliminated from the pupil measurements. With each blink, just before the eye is completely closed the pupil is partly obscured. In addition, eye closure gives rise to a change in pupil diameter because of the momentary variation in luminosity. Therefore, a period before and after each blink had to be removed from the data. To achieve this, blinks were identified and 200 ms before and 1000 ms after each blink were eliminated.

A third type of preprocessing already planned at design time (see Sect. 2) was to relate pupil diameter assessed during reading to the baseline value measured just before the reading of the text in question. In this way, long-term variations in pupil size that are not related to the reading tasks can be taken into account. But in fact, baseline values correlated strongly negatively (with a mean correlation of -0.72) with the reading pupil diameter obtained by subtracting the baseline from the raw data. Such high negative correlations indicate that baseline correction is not justified. Therefore, in a first step, raw pupil data was analyzed for each subject as well as across all participants.

Text effects on single subject level were tested with analyses of variance for independent measurements. Even though difficulty effects were significant for each subject, the difficult texts gave rise to larger pupil diameters for only 6 of them, whereas the opposite relation was observed for the remaining 7 (see Fig. 4 for an example). Accordingly, there were no significant difficulty effects across all subjects ($F_{(3,46)} = 1.14,\ p > 0.3$, see Fig. 5).

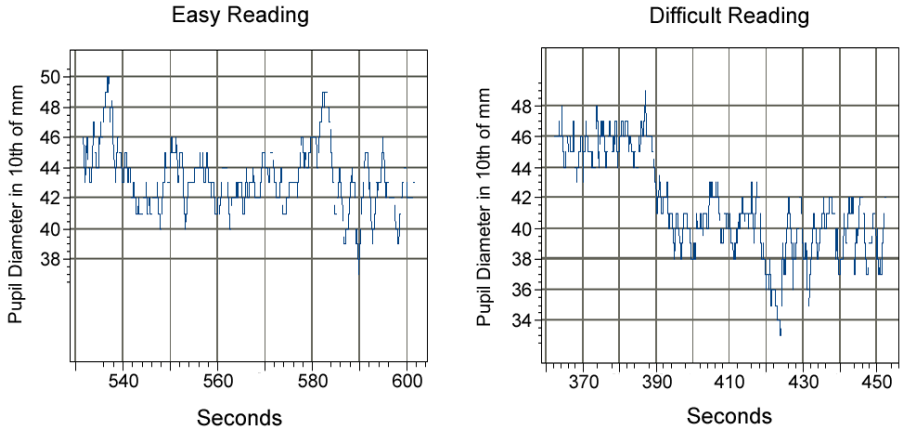

Fig. 4. Pupil diameter for one subject while reading an easy and a difficult text, respectively.

Because the lack of any difference in the pupil diameters was surprising, we conducted a number of additional tests, for example, considering only the first few seconds of the text reading; and correcting pupil diameter measurements to take into account differences in the measurements due to different points of gaze. In all analyses, there was no hint of a consistent difference in pupil diameters between the two conditions.

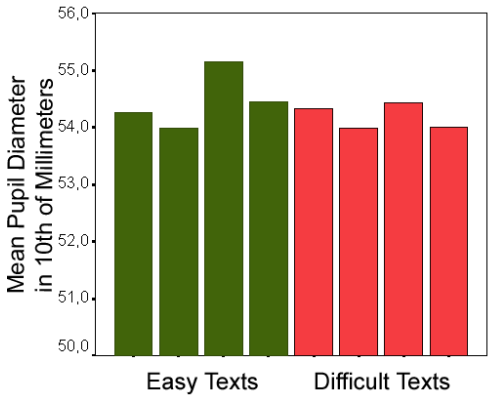

Fig. 5. Mean pupil diameter for all subjects during the reading of easy and difficult texts.

4 Discussion

Of the four measures of cognitive load used in this study, three—reading speed, subjective load and P300 amplitude—show a clear effect of text difficulty, indicating that demanding texts indeed induced an increased cognitive load. But this difference in load was not observable in pupil diameter. This result is surprising in that many previous studies (see Sect. 1.2) had consistently reported the sensitivity of the pupil size measure. But they did so in rather strictly controlled settings, and the present study suggests that their results do not generalize to settings that are typical of adaptive hypermedia systems.

This unexpected result has recently been confirmed by independent research: Iqbal et al. [12] examined pupil-size sensitivity to load variations in four different tasks, one of which was a reading task similar to the one employed in this experiment. For two of these tasks (file management on a computer and the reading of texts), no overall pupil size difference between easy and difficult conditions could be found. On the other hand, an analysis of the file management task on the subtask level revealed pupil size differences corresponding to the level of cognitive load in the subtasks. So it seems that pupil size may differ between easy and difficult conditions only in certain periods of a task. Whereas identification of appropriate subtasks was possible for the file management task, it is not obvious how a reasonable decomposition could be achieved for reading. Moreover, such a decomposition would most likely be dependent on the particular text. Consequently, our results and those of Iqbal et al. [12] indicate that pupil-size—although it may be sensitive to load in general—is not a suitable measure of load for tasks that involve continuous reading.

Although this result is a negative result, we believe that it is worth drawing attention to. There have been many reports of relationships between pupil diameter and cognitive load; and more generally, there has been a lot of optimism about the prospects of using physiological methods for the assessment of computer users' cognitive or affec-

tive states. If only positive results along these lines are published, a seriously distorted impression of the potential of these methods is likely to arise. Our study illustrates that the utility of physiological assessment methods can depend strongly on the nature of the task and the situation of use.

For the type of setting considered here, using behavioral indicators instead of physiological measures may be more appropriate. As was mentioned above, (Sect. 3) reading speed was considerably higher for easy texts. Consequently, reading speed might be used to assess the cognitive load of a user currently studying a hypertext page. Of course, one has to find a suitable way to assess speed. One possibility is to utilize the eye tracker to record the time taken to read a text of known length. The advantage of this approach would be that—as long as the user is reading—an up-to-date estimate of load is available. This particular approach can be realized only when information about the placement of text on the screen is available. But in other situations, it may be possible to assess reading speed on the basis of actions like button presses and mouse clicks.

References

1. Wilson, G.F., Eggemeier, F.T.: Psychophysiological assessment of workload in multi-task environments. In Damos, D.L., ed.: Multiple-Task Performance. Taylor and Francis, London (1991) 329 – 360
2. Tsang, P., Wilson, G.F.: Mental workload. In Salvendy, G., ed.: Handbook of Human Factors and Ergonomics. Wiley, New York (1997) 417 – 449
3. O'Donnell, R.D., Eggemeier, F.T.: Workload assessment methodology. In Boff, K., Kauffmann, L., Thomas, J., eds.: Handbook of Perception and Human Performance. Wiley, New York (1986) 42–1 – 42–49
4. Beatty, J.: Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin **91** (1982) 276 – 292
5. Kramer, A.F.: Physiological metrics of mental workload: A review of recent progress. In Damos, D.L., ed.: Multiple-Task Performance. Taylor and Francis, London (1991) 279 – 327
6. Hyönä, J., Tommola, J., Alaja, A.M.: Pupil dilation as a measure of processing load in simulaneous interprtation and other language tasks. The Quarterly Journal of Experimental Psychology **48A** (1995) 598 – 612
7. Just, M.A., Carpenter, P.A.: The intensity dimension of thought: Pupillometric indices of sentence processing. Canadian Journal of Experimental Psychology **47(2)** (1993) 310 – 339
8. Dionisio, D.P., Granholm, E., Hillix, W.A., Perrine, W.F.: Differentiation of deception using pupillary responses as an index of cognitive processing. Psychophysiology **38** (2001) 205 – 211
9. Sirevaag, E.J., Kramer, A.F., Coles, M.G., Donchin, E.: Resource reciprocity: An event-related brain potential analysis. Acta Psychologica **70** (1989) 77 – 97
10. Friedman, D., Cycowicz, Y.M., Gaeta, H.: The novelty P3: An event-related brain potential (ERP) sign of the brain's evaluation of novelty. Neuroscience and Biobehavioral Reviews **25(4)** (2001) 355 – 373
11. Ullsperger, P., Freude, G., Erdmann, U.: Auditory probe sensitivity to mental workload changes - an event-related potential study. International Journal of Psychophysiology **40** (2001) 201 – 209
12. Iqbal, S.T., Zheng, X.S., Bailey, B.P.: Task-evoked pupillary response to mental workload in human-computer interaction. In: Proceedings of the ACM Conference on Human Factors in Computing Systems, Vienna, Austria (2004) 1477 – 1480