Behavior Research Methods
2009, 41 (3), 868-875
doi:10.3758/BRM.41.3.868

Exploring workload and attention measurements with uLog mouse data

Annemiek van Drunen
Technical University Eindhoven, Eindhoven, The Netherlands

Egon L. van den Broek
University of Twente, Enschede, The Netherlands

and

Andrew J. Spink and Tobias Heffelaar
Noldus Information Technology, Wageningen, The Netherlands

User–system interactions (e.g., mouse clicks and movements) can be logged with the uLog computer program. A Web-based study with 20 participants was conducted to investigate the feasibility of using uLog data as an indicator of workload and attention. Eye fixation, heart rate variability (HRV), and skin conductance were used to unveil users' workload and attention and, hence, to validate uLog data as indicators of these. Results on one of the Tasks did indeed show correlations between uLog data and HRV. This is a promising first step toward the validation of uLog mouse data as indicators of workload and attention.

When developing a user interface (UI), it is essential to find the balance between (1) grasping and holding attention and (2) providing optimal information density in the center and periphery of focus. This balance is important in order to prevent information overload (Chun, 2000; Johnson & Proctor, 2004).

Logging user–system interactions helps in evaluating UIs (Alexander, Cockburn, & Lobb, 2008; Cooke, 2006; Kukreja, Stevenson, & Ritter, 2006; Nielsen, 1993; Trewin, 1998; Westerman et al., 1996). Throughout the years, *Behavior Research Methods* has published various programs that were able to log these interactions (e.g., Alexander et al., 2008; Kukreja et al., 2006; Trewin, 1998; Westerman et al., 1996).

In the present study, we use Noldus Information Technology's uLog (Noldus Information Technology, 2008) to record user–system interactions, including mouse clicks, mouse traces, and keyboard input. Such data can help to determine the usability of existing UIs and to develop guidelines for interface design, as is illustrated in Alexander et al. (2008), Cooke (2006), Kukreja et al. (2006), Nielsen (1993), Trewin (1998), and Westerman et al. (1996). For example, the parts of the UI that draw people's attention can be revealed. The aim of the present study is to investigate the possibility of measuring workload and attention with uLog's user–system interaction data.

In this article, we adopt the definition of *attention* provided by Smelser and Baltes (2001). They stated, "within information processing psychology, the term 'attention' refers to a mechanism that selects a spatially coherent subset of sensory information from among all information available" (p. 868). They also stated that "attention is a general term for selectivity in perception. The selectivity implies that at any instant a perceiving organism focuses on certain aspects of the stimulus situation to the exclusion of other aspects" (p. 878).

For our definition of *workload*, we refer to Chapters 41 and 42 of the *Handbook of Perception and Human Performance* (Boff, Kaufman, & Thomas, 1986). In Chapter 41, Gopher and Donchin (1986) defined mental workload as

> an attribute of the information processing and control systems that mediate between stimuli, rules, and responses. Mental workload is an attribute of the person–task loop, and the effects of the workload on human performance can therefore be examined only in relation to a model of human information processing. (p. 41.3)

In chapter 42, O'Donnell and Eggemeier (1986) included "that portion of the operator's limited capacity actually required to perform a particular task" (p. 42.2).

All mental processing requires some resources. The greater the task demand, the greater the processing resources needed to maintain performance at a certain level. If task demands exceed available resources, performance falters. Therefore, having some indications of the expected resources would be indispensable for ensuring optimal performance.

A. van Drunen, a.v.drunen@tue.nl

The collection of user–system interactions, such as mouse clicks and hovers, requires only that uLog (or similar software; see, e.g., Cegarra & Chevalier, 2008) be installed on the user's computer. This software and research method is inexpensive and totally nonintrusive, which could make uLog suitable for measuring psychological constructs, such as attention and workload. However, measuring these two constructs using user–system interactions requires thorough validation (e.g., Verwey & Veltman, 1996). For the present study, we chose to record three types of measurements suitable for such validation (cf. Cegarra & Chevalier, 2008): (1) psychophysiological measurements, which include electrocardiograms (ECGs; Berntson et al., 1997; Cacioppo, Tassinary, & Berntson, 2007) and electrodermal activity (EDA; Boucsein, 1992; Cacioppo et al., 2007); (2) eyetracking (Duchowski, 2002, 2007); and (3) thinking aloud (Nielsen, 1993; Van Someren, Barnard, & Sandberg, 1994).

Physiological measures are often used to support and interpret behavioral measures. The domain that studies the interaction between those physiological and behavioral measures is *psychophysiology* (Boccia & Roberts, 2000; Cacioppo et al., 2007; Fairclough, 2009). In the domain of psychophysiology, workload has been shown to be related to heart rate variability (HRV; i.e., the variability of heart rate over time; Berntson et al., 1997). HRV can be determined via ECG. In several studies, a negative correlation was found between HRV and workload (e.g., Hansen, Johnsen, & Thayer, 2003; Middleton, Sharma, Agouzoul, Sahakian, & Robbins, 1999; Rowe, Sibert, & Irwin, 1998; Wastell & Newman, 1996). However, other research has shown contradicting results on the correlation between the user's HRV and his or her mental state.

EDA can also be determined and is expressed through skin conductance (SC), which is the electrical resistance of the skin surface (Boucsein, 1992; Stern, Ray, & Quigley, 2001). Although they can be heavily influenced by factors like emotions and stress (see, e.g., van den Broek, Janssen, Westerink, & Healey, 2009), differences in SC can also indicate possible changes in attention or workload (see, e.g., Critchley, Elliott, Mathias, & Dolan, 2000).

An eyetracker detects saccades (i.e., rapid movements) and fixation points of the users' points of gaze. When exploring a Web site or reading a text, users typically show a pattern of saccades followed by a fixation (Duchowski, 2002, 2007). Such patterns are recorded by the eyetracker and are considered to indicate attention (Pashler, 1998). In addition, some hints are provided for possible correlations between location and number of eye fixations, as well as for those between mouse clicks and mouse hovers (see, e.g., Cooke, 2006). In general, the ability to record and replay both eye movements and mouse clicks over screens is extremely useful. Moreover, many problems of UIs can be determined through task-completion times and the amount of mouse behavior necessary to fulfill an assignment (Alexander et al., 2008; Cooke, 2006; Kukreja et al., 2006; Trewin, 1998; Westerman et al., 1996).

The think-aloud method asks users or participants to work on a task and verbalize their thoughts during the execution of that task (Nielsen, 1993; Van Someren et al., 1994). This method is common in cognitive and human–computer interaction research, although there is some discussion as to whether the method is a good one. One issue is the possible influence that the verbalization could have on other measurements or on task execution itself (see, e.g., Cooke & Cuddihy, 2005). Additionally, some practice is needed before people are capable of thinking aloud (Jaspers, Steen, van den Bos, & Geenen, 2004; Van Someren et al., 1994).

The combination of these three types of data is unique (cf. Goldberg, Stimson, Lewenstein, Scott, & Wichansky, 2002) and will supplement uLog data to validate its suitability for measuring attention and workload. This approach is better known as *triangulation*—that is, "the strategy of using multiple operationalizations of constructs to help separate the construct under consideration from other irrelevancies in the operationalization" (Smelser & Baltes, 2001, p. 15901). Triangulation reduces the amount of error in explaining users' behavior by isolating and detecting the constructs more easily.

In the following section, we will describe a pilot experiment that was conducted to reveal possible interaction influences of the different measurements just introduced. After that, the main experiment will be introduced. For both experiments, the methods and results are denoted. The General Discussion section lists the pros and cons of the study, describes the conclusions, and sums up recommendations for future research.

PILOT EXPERIMENT

Method

Participants. Six colleagues from Noldus IT participated in the experiment. All of the participants cooperated on a voluntary basis during working hours. The participants were selected using three criteria: They had to be unfamiliar with uLog, eyetracking, and physiological data acquisition. There were 2 female and 4 male participants ranging from 22 to 44 years of age ($M = 35.8$ years).

Apparatus and Stimulus. The experiment took place in Noldus's experience lab, which consists of an experimentation room and a control room (Figure 1). During the experiment, everything was monitored

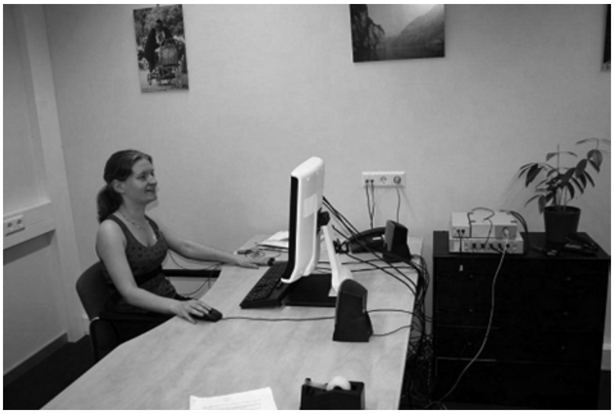

Figure 1. The experimental setup, where a participant is connected to the apparatus and sitting behind the eyetracking monitor.

from the control room. In the pilot test, the verbalized thoughts of the participants were recorded with the use of video recording software.

At the beginning and end of the experiment, the participants had to fill out a questionnaire, which gathered data concerning the following information: (1) demographics, including age, gender, and handedness; (2) computer usage, including mouse hand, Internet usage, and Web page development; (3) knowledge of Noldus IT, including knowledge of Noldus or of Noldus's Web site; (4) experimental experience—that is, experience with taking part in or doing research; and (5) feelings experienced, such as *happy*, *tensed*, *tired*, and *confident*.

The introduction of the pretest questionnaire included information about the outline of the experiment. At the end of the experiment, a second questionnaire was filled out. This contained the same questions about feelings. In addition to that, the experience with the experiment, with the tasks, and with the apparatus were evaluated.

Four search Tasks on the same Internet page were performed by the participants. The page consisted of 24 color pictures of cartoon faces. The pictures consisted of combinations of a certain hair color (i.e., blond, gray, black, or brown), eye color (i.e., blue, green, or brown), and facial expression (i.e., smiling or sad). The same face was used, except for the manipulated feature. To complete each task, the participants had to click as fast as possible on the cartoon face that met the task description that was displayed on the introduction screen of each task.

During the tasks, the participants' ECG, EDA, eye movements, and eye fixations were measured, and the verbalized thoughts of the participants were recorded. Details concerning the hardware, software, and settings used will be explained in the "Data and data reduction" section below.

Procedure. The pilot experiment consisted of four conditions for each of the participants. Each condition comprised the Tasks described immediately above. In each condition, a different combination of measurements was taken and another (comparable) task had to be carried out. The four conditions encompassed the following measurement combinations, which were given in a different order to each participant: (1) eyetracking, physiological, and thinking aloud; (2) eyetracking and thinking aloud; (3) eyetracking and physiological; and (4) eyetracking (see Figure 2).

The experiment comprised eight phases. Each phase is denoted, indicating the moment of the experiment to which they apply. The phases that took place at the beginning of the experiment included the following:

1. The participant was asked to fill out the first questionnaire.

2. A short explanation of the experiment was given by the experimenter.

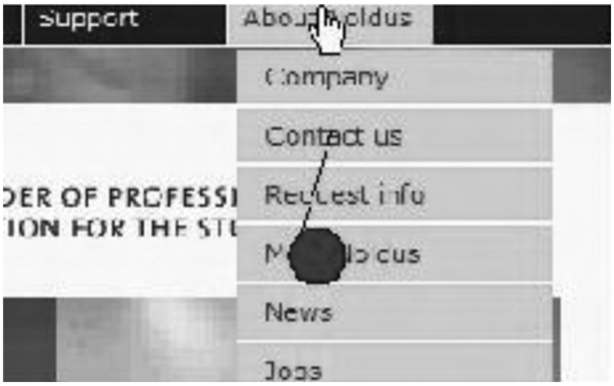

Figure 2. A section of the stimulus Web site with a drop-down menu. The dot indicates a fixation point displayed with the eyetracker software; the white hand is the mouse pointer.

3. The participant practiced thinking aloud by verbalizing their thoughts while putting together the parts of a 3-D block puzzle (ThinkFun, Inc., 2003).

The following steps were repeated for each of the four conditions.

4. If required for the condition, the electrodes for the physiological measurement were attached, and the apparatus and software were calibrated.

5. The eyetracker was calibrated.

During the execution of each condition, the experiment leader observed the participants and coordinated the Tasks from an experimentation room. Between tasks, the researcher came to the experimentation room to connect or disconnect the appropriate apparatus for the next condition. Then the following occurred.

6. If required for the condition, a cooling-down period of 5 min was included, during which the baseline measure for the physiological data was taken.

7. The condition was executed.

8. At the end of the experiment, the participant completed the questionnaire about the experiment and was thanked.

In total, the experiment took approximately 75 min for each participant.

Data and data reduction. The recorded data were imported into The Observer XT 8.0 (Zimmerman, Bolhuis, Willemsen, Meyer, & Noldus, in press), synchronized, and exported to text format. (The Observer XT is a software package that is widely used for the collection and synchronization of observational data.) For every participant, each of the Tasks and the cooling-down period was exported separately. Before analysis, preprocessing of the data took place. Explanations of the preprocessing and analysis procedure, as well as of the hardware, software, and the software's settings, are provided below.

Eyetracking data were gathered with a Tobii T60 eyetracker and ClearView software (Version 2.7.1). Fixations longer than 80 msec were extracted from the eyetracking data for each participant (Duchowski, 2002, 2007).

Physiological data were measured with an ADI Powerlab 8/30 ML 870 physiological data acquisition system, including the appropriate wires, connectors, and Chart software (Version 5.4.4) for ECG and SC measurements, both at a sample frequency of 2000 Hz. Visualizations of samples of the physiological data as measured by a participant are shown in Figure 3. The placement of the electrodes for ECG was carried out according to the modified Lead II placement (Stern et al., 2001). For the SC measurements, an SC strap was placed around the tops of the index and middle fingers of the hand that was not used for operating the mouse. Before strap placement, both the participant's hand and the strap were cleaned (Stern et al., 2001).

ADI Powerlab's predefined settings were used in the software for the ECG and EDA measurements, including a 50-Hz mains filter. This filter is integrated with the Chart software and used to clean the signal of noise and alias artifacts (for further details, see ADInstruments, 2007). Before the recording of the experiment started, we checked whether the software was receiving correct signals, and electrodes were replaced or changed if necessary. After the experiment, for all participants, the R-peaks of the ECG signal were detected, and, subsequently, the R–R interval was calculated. As a measure of HRV in the time domain, the standard deviation of the cardiac cycle was calculated from the R–R interval. Respiration data were gathered to control for artifacts in the ECG data because breathing can influence the ECG signal. All of the physiological data were analyzed for differences within and between the different Tasks and participants. To prepare the physiological data for these analyses, all of the streams were imported into The Observer XT and synchronized with the other data.

The data from the video files were used to analyze the participants' verbalized thoughts and to divide the data into different parts. For each task and cooling-down period, a separate file was generated. These video files and observational logs were recorded and synchronized with the other data with the use of The Observer XT.

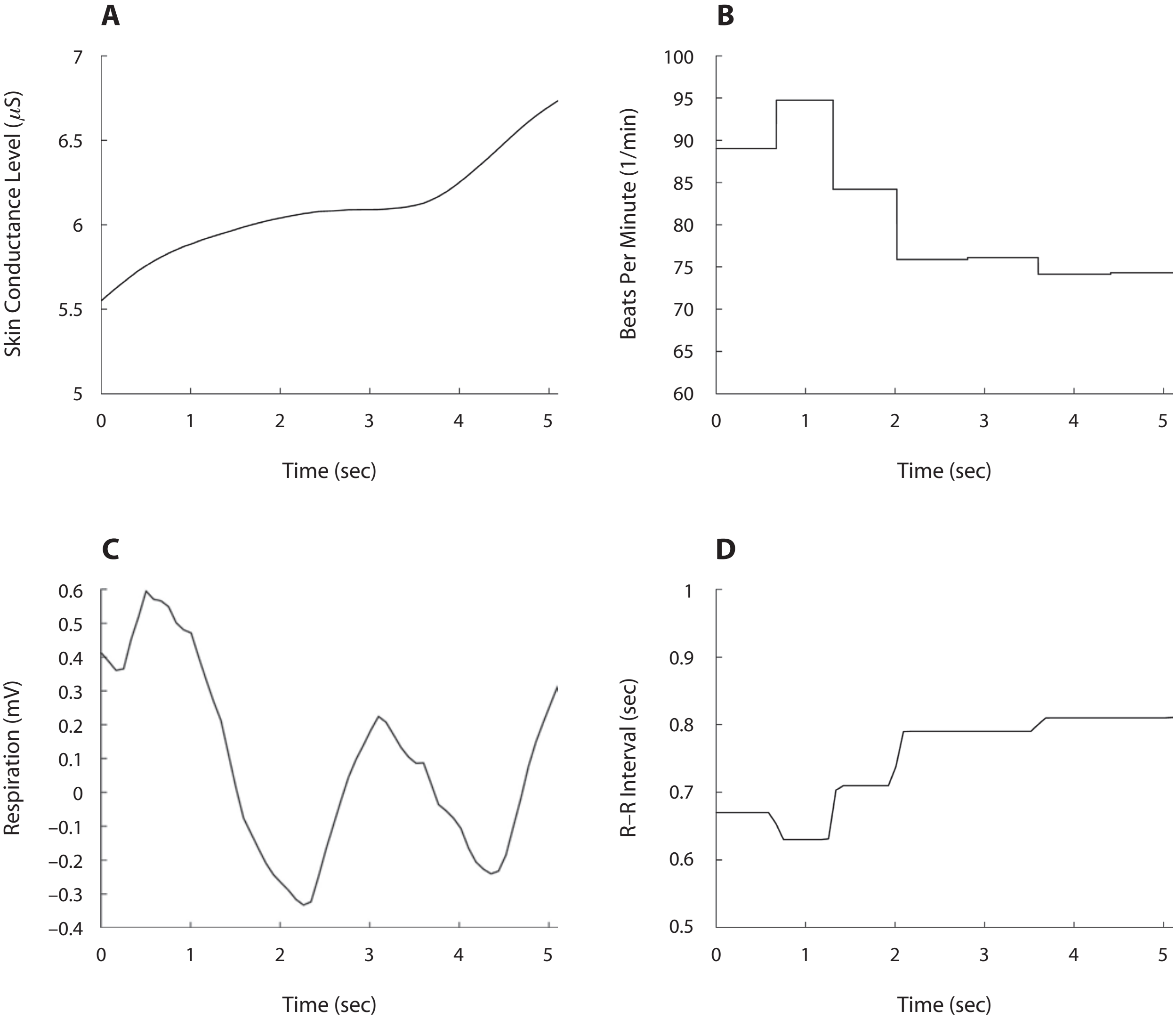

Figure 3. Visualization of the physiological measures. (A) Skin conductance level. (B) Heart rate in beats per minute. (C) The respiration signal in mV. (D) The R–R interval of the electrocardiogram. These physiological measures are determined in parallel, from 1 participant, during the execution of the tasks. This visualization shows the same 5 sec of data for 1 participant, from the different measurements.

Results and Discussion

The data of the pilot experiment showed that there was an effect of the think-aloud condition on both the physiological and eyetracker data. In addition, all 6 participants stated that, while carrying out the Tasks of the experiment, they found thinking aloud to be hard (and sometimes even impossible). This effect has been shown in other research (e.g., Cooke & Cuddihy, 2005; Van Someren et al., 1994).

As Nielsen (1993, p. 196) has stated,

> Thinking aloud seems very unnatural to most people, and some test users have great difficulties in keeping up a steady stream of utterances as they use the system. Not only can the unnaturalness of the thinking aloud simulation make the test harder to conduct, but it can also impact their results. First, the need to verbalize can slow users down, thus making any performance measurements less representative of the users' regular working speed. Second, users' problem solving behavior can be influenced by the very fact that they are verbalizing their thoughts.

Taken together, the results of this pilot study confirm Nielsen's (1993) statement through the participants' experience, the psychophysiological recordings, the eyetracker data, and the performance measures. This result stresses the downside of using the think-aloud protocol. Consequently, for the main experiment, the think-aloud measurement was left out. For the other measurements, no interactions were found; the rest of the setup was suitable, however, and was therefore used in the main experiment.

MAIN EXPERIMENT

Method

The procedure, apparatus, and software of the pilot experiment and the main experiment were similar, except for the following differences.

Participants. Complete data of 14 participants were analyzed. Of the participants, 14 were male, and 10 were female. Their mean age was 30.33 years, ranging from 18 to 54 years. Only right-handed people participated in the experiment, but 1 participant used her left hand to operate the mouse. Before the participants could take part in the experiment, they had to read and sign an informed consent document that also explained the outline of the experiment. The participants were recruited at Wageningen University, the public library, and other public places in Wageningen, The Netherlands. They were paid for their participation.

Materials. The participants carried out four predefined Tasks using the Noldus IT Web site. During each task, the participants were asked to search for a certain item or part of the Web site, using task descriptions. During the fifth task, the participants had to explore for the Web site for about 3 min. They were asked to answer some questions about the Web site at the end of the task. This was done to ensure that the participants stayed focused on the content of the Web site; the participants had been told before the experiment that they would have to perform this fifth task at the end.

Procedure. The experimental procedure was the same for the pilot test, except for the following four aspects: (1) In the main experiment, all measurements were taken for all tasks; (2) uLog was used to measure and record the user–system interactions of the participants, such as mouse clicks and hovers; (3) the order of the four predefined Tasks was randomized among participants, except for the fifth “free surfing” task, which was consistently the final task; and (4) unlike in the pilot test, thinking aloud was omitted.

Data and data reduction. The data and the preprocessing of data were the same as for the pilot test, except for the different aspects and data mentioned above.

In the experiment, an experimental stand-alone version of uLog was used, which generated XML files (World Wide Web Consortium, n.d.). All of the other measurements were directed to and controlled from within The Observer XT software. After the experiments, the uLog data were imported into Observer XT to synchronize them with the other data sources. A summary of the data that were taken from every data source can be found in Table 1.

Table 1
The Data Sources Gathered in the Experiment and Their Accompanying Features

Source	Feature
Cardiovascular activity	Mean BPM Mean HRV and *SD*
Electrodermal activity	Mean SC and *SD*
Postquestionnaire	Perceived task difficulty
Pre- and postquestionnaire	Feelings Demographic data
Eyetracking data	Number of fixations
uLog data	Number of mouse clicks Number of hovers Total mouse behavior (clicks + hovers)
Video files and logs	Task duration

Note—BPM, heart rate in beats per minute, as determined via electrocardiogram; HRV, heart rate variability, in milliseconds; SC, skin conductance, in μS; *SD*, standard deviation.

Results

Questionnaires. The pre- and postquestionnaire contained the same four statements about the participants’ feelings, which the participants had to either agree or disagree with, using a 5-point Likert scale. Each answer was given a score from 1 (*totally disagree*) to 5 (*totally agree*). The ordinal and nonparametric questionnaire data were analyzed with a Spearman rho correlation and a Wilcoxon signed-rank test. The goal of these analyses was to determine whether the experiment affected participants’ feelings. The results of these analyses are shown in Table 2. With the Wilcoxon signed-rank test, the difference between paired ordinal variables can be analyzed by ranking and summing up the differences of the scores within each of these pairs (Higgins, 2004). The results denote the existence of either a positive or a negative difference between the pre- and postquestionnaire responses of any individual participant. These differences are referred to as *positive ranks* and *negative ranks*, respectively. For each participant, each couple of equal pretest and posttest scores on a question was counted as a tie. These analyses were carried out with SPSS.

Significant correlations were found between the pre- and postquestionnaire results. The results of the Wilcoxon signed-rank test gave insight into the differences between the pre- and postquestionnaire pairs for each feeling of each participant. For the question on happiness, there were two cases involving higher happiness scores on the posttest than on the pretest questionnaire, and each of those two cases had a positive rank of 3. For the other 12 participants, the scores remained the same. For the question about tension, there was one case involving a higher score on the posttest than on the pretest questionnaire, giving a sum of positive ranks of 3. Furthermore, this question had four cases with a higher score on tension after the experiment, with a sum of ranks of 12. In the remaining nine cases, there was no difference found between the pre- and postquestionnaires. For the question about confidence, there were three cases having a higher score on the postquestionnaire, with a sum of ranks of 10, and there were two cases having a higher score on the prequestionnaire, with a sum of ranks of 5. In the remaining nine cases, no difference in scores was found between the pre- and postquestionnaires. The last question concerning participants’ feelings assessed the tiredness of

Table 2
Analyses Results of a Comparison Between the Pre- and Postexperimental Questions

Feeling	r_s	p	Wilcoxon Positive Ranks	Wilcoxon Negative Ranks	Wilcoxon Ties	Sum of Positive Ranks	Sum of Negative Ranks
Happy 2–Happy 1	1.000	.000	2	0	12	3	0
Tensed 2–Tensed 1	.886	.000	1	4	9	3	12
Confident 2–Confident 1	.770	.006	3	2	9	10	5
Tired 2–Tired 1	.851	.001	4	2	8	12.5	8.5

the participants. In eight cases, no difference was found between the pre- and postquestionnaire scores. Four cases had a higher score on the postquestionnaire, with a sum of ranks of 12.5; and in two cases, the score on the prequestionnaire was higher, with a sum of ranks of 8.5.

Because of time limitations, only two subtasks—Task 1 and Task 2—were analyzed. The data from these Tasks were also combined in a data set (combined task) and analyzed as a whole. In total, four sets of analyses were done: a general analysis (i.e., Tasks 1 and 2), Task 1 and Task 2 analyzed separately, and, finally, a cross-task analysis comparing the results of the two Tasks together.

Analysis. To complete Tasks 1 and 2, the participants needed on average 12.32 sec (*SD* = 3.20). Negative correlations were found between mean HRV and perceived task difficulty ($r = -.666$, $p = .025$) and between total mouse behavior and duration ($r = -.641$, $p = .046$). The measure for task difficulty was derived from a question in the postquestionnaire on which participants had to answer using the aforementioned Likert scale. *Total mouse behavior* was defined as the sum of the total number of clicks and hovers. A positive correlation was found between total mouse behavior and fixation ($r = .789$, $p = .007$). Negative correlations were found between duration and both mean HRV ($r = -.792$, $p = .004$) and HRV *SD* ($r = -.843$, $p = .001$). In this task, the number of fixations correlated positively with total mouse behavior ($r = .789$, $p = .007$), mouse hovers ($r = .772$, $p = .009$), and mouse clicks ($r = .661$, $p = .037$). Positive correlations were found between mean BPM and both SC ($r = .683$, $p = .021$) and SC *SD* ($r = .652$, $p = .030$).

The mean time participants needed to complete Task 1 was 6.00 sec (*SD* = 1.96). As mentioned in the Method section, we decided to normalize the mouse clicks, mouse hovers, total mouse behavior, and fixations over the task duration for every participant individually and for all Tasks analyzed. The *SD* of the HRV correlated significantly with total mouse behavior ($r = .700$, $p = .010$) and number of mouse hovers ($r = .710$, $p = .010$). The mean HRV also correlated significantly with both mouse hovers ($r = .814$, $p < .001$) and total mouse behavior ($r = .787$, $p = .002$). The latter correlation is displayed in Figure 4; the dots indicate the different values for each participant, and the R^2 value displays the regression coefficient. Significant negative correlations were found between duration and both total mouse behavior ($r = -.834$, $p < .001$) and mouse hovers ($r = -.816$, $p = .001$). The number of mouse clicks correlated with both SC ($r = .726$, $p = .008$) and BPM ($r = .596$, $p = .050$).

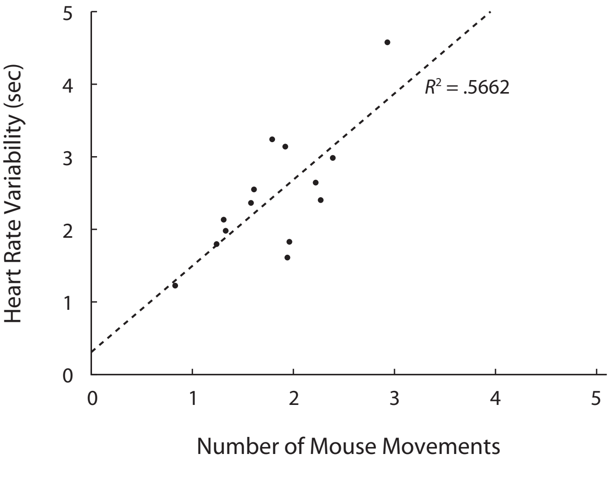

Figure 4. Correlation between mean heart rate variability and total mouse behavior of Task 1 (the sum of the number of mouse clicks and hovers). Each dot is a data point of 1 participant. The straight line displays the trend line, and the R^2 value displays the regression coefficient.

The mean time that participants needed to complete Task 2 was 5.83 sec (*SD* = 2.02). To compare the different data sources, correlation analyses were carried out. A negative correlation was found between mean HRV and perceived task difficulty (TaskHard) ($r = -.684$, $p = .020$). Further correlations were found between mean BPM and both mean SC ($r = .731$, $p = .011$) and SC *SD* ($r = .682$, $p = .021$).

To compare the results of Task 1 and Task 2, the data were correlated. A negative correlation was found between fixations of Task 1 and Task 2 ($r = -.646$, $p = .017$). A positive correlation was found between HRV *SD* ($r = .581$, $p = .037$) and the data of the two tasks.

GENERAL DISCUSSION

The results of the main experiment showed high correlations between the preexperimental and postexperimental questionnaire results. These correlations showed that, overall, participants' stated feelings (e.g., happiness, tiredness, and confidence) were not changed due to the experiment. This was an important control since people's feelings are known to heavily influence experimental results—especially physiological signals (van den Broek et al., 2009). When focusing on the individual level, we found the means and variances of the differences between pre- and postquestionnaire responses to be low, indicating rather stable data.

When analyzing Task 1, we found significant correlations between mouse hovers and total mouse behavior and the mean HRV parameters and *SD*s. This is in line with various other studies in which low levels of HRV indicated a high workload and vice versa (Middleton et al., 1999; Rowe et al., 1998; van der Molen, Boomsma, Jennings, & Nieuwboer, 1989). Therefore, it can be concluded that user–system interaction data can be used to indicate perceived workload. However, this conclusion needs further validation because HRV is found to be a sensitive measure. For example, no consensus exists on which aspects of the HRV signal are most suitable for measuring workload and attention (Berntson et al., 1997). According to Berntson et al., a minimal measurement duration of 2 min is required in order to get reliable results from HRV. Although the duration of the experimental task was more than 2 min, the signals we analyzed were shorter than that.

Tasks 1 and 2 seemed to be comparable: Participants had to go to the menu bar of the Web site and select a certain menu item from the drop-down list. However, the correlations found between HRV and user–system interaction data in Task 1 were not found in Task 2. A possible cause could be differences in task execution and, consequently, a deviation in the patterns of mouse clicks and mouse movements. At first glance, no obvious differences in task execution were visible during manual analysis of the experimental video data, and the Tasks and related description were well defined. However, the realistic nature of the Tasks could explain deviations in task execution. The choice for realistic Tasks was made because the aim of programs like uLog is to measure user–system behavior in a realistic setting. A major concern of in vivo experimental settings is a lower level of control. Such settings increase external validity, but they decrease internal validity (for recent reviews, see Fairclough, 2009; van den Broek et al., 2009).

This study can be seen as a first step toward nonintrusive recording of users' workload and attention with user–system interaction mouse data. uLog can record user–system interaction without user awareness and without any disturbance or delay in the interaction process. However, uLog data on its own cannot provide enough information and therefore requires a selection of other, additional measures. The present experimental setting was rather intrusive, which is undesirable in practice.

A balance is needed between optimally complementing uLog recordings and limiting intrusiveness of the recordings for the user. This research is an explorative step toward the measurement of psychological constructs using user–system interaction data—specifically, mouse movements and clicks. Future research could (1) investigate the optimal trade-off between measurements; (2) strive to validate the correlation between mouse behavior and HRV data; and (3) focus on better controlled, in vitro, stimuli, in order to diminish intersubject differences. One of the few known examples of this type of research is that of Tuch, Bargas-Avila, Opwis, and Wilhelm (in press). In their study, well-controlled stimuli were used for evaluating the visual complexity of Web sites and its influence on physiological measures.

Whereas our research clearly has its limitations, it illustrates the complexity of unraveling cognitive constructs using psychophysiological signals. It also illustrates the limitations of the think-aloud protocol for experimental research. Moreover, through its approach of triangulation, some useful indications for the interpretation of mouse behavior in terms of cognitive constructs are found. In this line, a vast amount of additional research is needed in order to further explore and strengthen the findings of the present study. Nevertheless, this research could be a significant step toward a better understanding of the relation between users' mental states and their behavior.

AUTHOR NOTE

The authors thank the reviewers for their useful comments and suggestions. The experiments described in this article were carried out at Noldus IT, Wageningen, The Netherlands. Without the help of many colleagues there, the experiments would not have been possible; therefore, the authors thank all of the people at Noldus who helped during the several stages of this research. Finally, we thank all of the test participants and interviewees. Correspondence concerning this article should be addressed to A. van Drunen, Industrial Design, Technical University Eindhoven, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (e-mail: a.v.drunen@tue.nl).

REFERENCES

ADInstruments (2007). Chart software manual (Version 5.5) [Software]. Bella Vista, Australia: ADInstruments Pty Ltd.

Alexander, J., Cockburn, A., & Lobb, R. (2008). AppMonitor: A tool for recording user actions in unmodified Windows applications. *Behavior Research Methods*, **40**, 413-421. doi:10.3758/BRM.40.2.413

Berntson, G. G., Bigger, T. J., Jr., Eckberg, D. L., Grossman, P., Kaufmann, P. G., Malik, M., et al. (1997). Heart rate variability: Origins, methods, and interpretive caveats. *Psychophysiology*, **34**, 623-648. doi:10.1111/j.1469-8986.1997.tb02140.x

Boccia, M. L., & Roberts, J. E. (2000). Computer-assisted integration of physiological and behavioral measures. In T. Thompson, D. Felce, & F. J. Symons (Eds.), *Behavioral observation: Technology and applications in developmental disabilities* (pp. 83-97). Baltimore: Brookes.

Boff, K. R., Kaufman, L., & Thomas, J. P. (Eds.) (1986). *Handbook of perception and human performance*. New York: Wiley.

Boucsein, W. (1992). *Electrodermal activity*. New York: Plenum.

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2007). *Handbook of psychophysiology* (3rd ed.). New York: Cambridge University Press.

Cegarra, J., & Chevalier, A. (2008). The use of Tholos software for combining measures of mental workload: Toward theoretical and methodological improvements. *Behavior Research Methods*, **40**, 988-1000. doi:10.3758/BRM.40.4.988

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, **4**, 170-178. doi:10.1016/S1364-6613(00)01476-5

Cooke, L. (2006). Is the mouse a "poor man's eye tracker"? In *Proceedings of the 53rd Annual Conference of the Society for Technical Communication* (pp. 252-255). Fairfax, VA: STC.

Cooke, L., & Cuddihy, E. (2005). Using eye tracking to address limitations in think-aloud protocol. In H. Grady (Ed.), *IEEE International Professional Communication Conference Proceedings* (pp. 653-658). Piscataway, NJ: IEEE. doi:10.1109/IPCC.2005.1494236

Critchley, H. D., Elliott, R., Mathias, C. J., & Dolan, R. J. (2000). Neural activity relating to generation and representation of galvanic skin conductance responses: A functional magnetic resonance imaging study. *Journal of Neuroscience*, **20**, 3033-3040.

Duchowski, A. T. (2002). A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers*, **34**, 455-470.

Duchowski, A. T. (2007). *Eye tracking methodology: Theory and practice* (2nd ed.). London: Springer.

Fairclough, S. H. (2009). Fundamentals of physiological computing. *Interacting With Computers*, **21**, 133-145. doi:10.1016/j.intcom.2008.10.011

Goldberg, J. H., Stimson, M. J., Lewenstein, M., Scott, N., & Wichansky, A. M. (2002). Eye tracking in Web search tasks: Design implications. In A. T. Duchowski, R. Vertegaal, & J. W. Senders (Eds.), *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications* (pp. 51-58). New York: ACM. doi:10.1145/507072.507082

Gopher, D., & Donchin, E. (1986). Workload: An examination of the concept. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance: Vol. II. Cognitive processes and performance* (pp. 41.1-41.49). New York: Wiley.

Hansen, A. L., Johnsen, B. H., & Thayer, J. F. (2003). Vagal influence on working memory and attention. *International Journal of Psychophysiology*, **48**, 263-274. doi:10.1016/S0167-8760(03)00073-4

Higgins, J. J. (2004). *Introduction to modern nonparametric statistics*. Stamford, CT: Duxbury.

Jaspers, M. W. M., Steen, T., van den Bos, C., & Geenen, M. (2004). The think aloud method: A guide to user interface design. *International Journal of Medical Informatics*, **73**, 781-795. doi:10.1016/j.ijmedinf.2004.08.003

Johnson, A., & Proctor, R. W. (2004). *Attention: Theory and practice*. London: Sage.

Kukreja, U., Stevenson, W. E., & Ritter, F. E. (2006). RUI: Recording user input from interfaces under Windows and Mac OS X. *Behavior Research Methods*, **38**, 656-659.

Middleton, H. C., Sharma, A., Agouzoul, D., Sahakian, B. J., & Robbins, T. W. (1999). Contrasts between the cardiovascular concomitants of tests of planning and attention. *Psychophysiology*, **36**, 610-618. doi:10.1111/1469-8986.3650610

Nielsen, J. (1993). *Usability engineering*. San Diego: Academic Press.

Noldus Information Technology (2008). uLog (Version 3.0) [Computer software]. Wageningen, The Netherlands: Author.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff, L. Kaufman, and J. P. Thomas (Eds.), *Handbook of perception and human performance: Vol. II. Cognitive processes and performance* (pp. 42.1-42.49). New York: Wiley.

Pashler, H. (Ed.) (1998). *Attention*. East Sussex, U.K.: Psychology Press.

Rowe, D. W., Sibert, J., & Irwin, D. (1998). Heart rate variability: Indicator of user state as an aid to human–computer interaction. In C.-M. Karat, A. Lund, J. Coutaz, & J. Karat (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 480-487). New York: ACM. doi:10.1145/274644.274709

Smelser, N. J., & Baltes, P. B. (Eds.) (2001). *International encyclopedia of the social & behavioral sciences*. Oxford: Elsevier.

Stern, R. M., Ray, W. J., & Quigley, K. S. (2001). *Psychophysiological recording* (2nd ed.). New York: Oxford University Press.

ThinkFun, Inc. (2003). *Block by block creative building game!* [Physical game]. Alexandria, VA: Author.

Trewin, S. (1998). InputLogger: General-purpose logging of keyboard and mouse events on an Apple Macintosh. *Behavior Research Methods, Instruments, & Computers*, **30**, 327-331.

Tuch, A. N., Bargas-Avila, J. A., Opwis, K., & Wilhelm, F. H. (in press). Visual complexity of Websites: Effects on users' experience, physiology, performance, and memory. *International Journal of Human–Computer Studies*. doi:10.1016/j.ijhcs.2009.04.002

van den Broek, E. L., Janssen, J. H., Westerink, J. H. D. M., & Healey, J. A. (2009). Prerequisites for affective signal processing (ASP). In P. Encarnação & A. Veloso (Eds.), *Biosignals 2009: Proceedings of the International Conference on Bio-Inspired Systems & Signal Processing* (pp. 426-433). Porto, Portugal: INSTICC.

van der Molen, M. W., Boomsma, D. I., Jennings, J. R., & Nieuwboer, R. T. (1989). Does the heart know what the eye sees? A cardiac/pupillometric analysis of motor preparation and response execution. *Psychophysiology*, **26**, 70-80. doi:10.1111/j.1469-8986.1989.tb03134.x

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modelling cognitive processes*. London: Academic Press.

Verwey, W. B., & Veltman, H. A. (1996). Detecting short periods of elevated workload: A comparison of nine workload assessment techniques. *Journal of Experimental Psychology: Applied*, **2**, 270-285.

Wastell, D. G., & Newman, M. (1996). Stress, control and computer system design: A psychophysiological field study. *Behaviour & Information Technology*, **15**, 183-192. doi:10.1080/014492996120247

Westerman, S. J., Hambly, S., Alder, C., Wyatt-Millington, C. W., Shryane, N. M., Crawshaw, C. M., & Hockey, G. R. J. (1996). Investigating the human–computer interface using the Datalogger. *Behavior Research Methods, Instruments, & Computers*, **28**, 603-606.

World Wide Web Consortium (n.d.). *Extensible Markup Language (XML)*. Retrieved May 17, 2009, from www.w3.org/XML/.

Zimmerman, P. H., Bolhuis, J. E., Willemsen, A., Meyer, E. S., & Noldus, P. J. J. (in press). The Observer XT: A tool for the integration and synchronization of multimodal signals. *Behavior Research Methods*.

(Manuscript received November 11, 2008;
revision accepted for publication April 21, 2009.)