Towards ex ante Prediction of User Performance: A novel NeuroIS Methodology based on Real-Time Measurement of Mental Effort

Ricardo Buettner
Inst. of Manag. & Inf. Syst. (mis)
FOM University Germany
ricardo.buettner@fom.de

Sebastian Sauer
mis at FOM University
Germany, Samueli Institute
sebastian.sauer@fom.de

Christian Maier
University of Bamberg
christian.maier@uni-bamberg.de

Andreas Eckhardt
Goethe University Frankfurt, eckhardt@wiwi.uni-frankfurt.de

Abstract

We propose a methodology of an ex ante prediction of users’ performance based on analyzing the pupillary diameter variability captured by ordinary eye-tracking systems. Based on a realistic large-scale experimental evaluation of our methodology we show promising results that pave the way for a dynamic real-time adaption of IT to the user’s mental effort and the expected user performance. Our non-invasive contact-free methodology can be applied cost-efficiently both in research and practical environments, without disturbing the participant/user.

1. Introduction

Achieving higher user performance in the use of information technology (IT) has for decades been a long-standing problem within information systems (IS) research [1]. For example, cognitive load theory (cf. [2]) conceptualised mental effort as an important IS construct that strongly influences user performance. Also, technology acceptance literature shows the high relevance of mental effort concerning perceived ease of use (e.g. [3, p. 325], [4, p. 697]). From various theoretical perspectives (e.g. cognitive load, task technology fit, job demands-resources) the adaptive regulation of a users’ mental effort is a critical requirement for user performance. But the means of measuring mental effort based on a user’s own subjective rating is not adequate, and in real-time such a measurement is impossible.

However, in recent years very interesting results have emerged from a new field called NeuroIS in which efforts have been made to determine a user’s mental effort based on objective psychophysiological measurements [5]–[7]. These results are very promising and possibly open the door to dynamic real-time adaption of IT to users’ mental efforts and in doing so could address one of the most challenging problems in IS research. In more detail, Dimoka et al. emphasised that “neurophysiological tools could assist in the design of metrics for complex constructs such as ... cognitive effort” [7, p.687] and “cognitive effort... can be reduced by effective IT designs” [6, p. 7].

However, despite these very promising (Neuro-)IS results we still have to establish research needs that contribute to the ambitious visions of the NeuroIS community because all existing research on predicting user performance based on real-time mental effort measurement is either (very) expensive, invasive or not contact-free. “There is a need for a direct measurement of information and cognitive overload, and neurophysiological tools have the potential to offer such a direct measurement.” [7, p. 685].

We aim to contribute to IS research by being the first to propose a method of ex ante prediction of user performance based on a non-invasive contact-free cost-efficient real-time measurement of mental effort. Mental effort is in all probability closely related to pivotal IS constructs such as perceived ease of use (cf. [3, p. 325], [4, p. 697], [6, p. 12], [8]). While our method can be applied cost-efficiently to research and practical environments without disturbing the participant/user our impact on IS research is very high. Our findings offer both researchers and industrial managers significant new knowledge that may pave the way towards a deeper understanding of user performance and thus towards the development of more advanced user performance management systems. Moreover, our method is based on a psychophysiological measurement instead of self-rated subjective participant/user reports and as such address an important IS measurement problem [6, p. 7]. In contrast to other physiological measurements such as electroencephalography (EEG), electrodermal activity (EDA), heart-rate variability (HRV), facial
electromyography (fEMG), functional magnetic resonance imaging (fMRI), positron emission tomography (PET), or magnetoencephalography (MEG) our method does not need a fixing of a cap/net (EEG) and/or electrodes (EDA/HRV/fEMG) or the use of a cost-intensive scanner (fMRI/PET/MEG). In addition, our method allows us to capture much more mental effort data points than any participant/user could ever be able to self-report (up to 600 Hz by using modern eye-tracking technology).

This paper is organized as follows: Section 2 spans the research background from the role of mental effort in IS research (2.1) via the mechanism on mental effort-based pupillary dilation (2.2) and the role of pupillary responses as a marker of mental effort in psychological research (2.3), to recent work on mental effort-based pupillary responses in IS-research – identifying the research need. In section 3 the research methodology is presented, including the application of the NeuroIS guidelines (3.1), the research model and the hypothesizing (3.2), the laboratory setting and measurement (3.3), the description of the test procedure (3.4), the sampling strategy (3.5), the data cleansing (3.6) as well as the description of the random forest method (3.7). Section 4 contains the results, including sample characteristics (4.1), the pupil diameter time series (4.2), the relationships between pupillary diameter variability and task difficulty (4.3) as well as user performance (4.4), predictive model results (4.5) and the model quality evaluation outcomes (4.6). The results are discussed in section 5. Finally, the conclusion is given in section 6, including research limitations (6.1) and future research needs (6.2).

2. Research Background

2.1. Mental effort in IS research

Determining a user’s mental effort is often mentioned as a fundamental problem in IS research (e.g. [9,10]), particularly in NeuroIS (e.g. [5]–[8,11]). It is remarkable that IS scholars have traditionally investigated a user’s mental effort and its derivatives [12] primarily based on user-perceived/non-objective measures (e.g. [13]–[16]) or even discussed the need for user workload measurements without any measurement proposal (e.g. [17]). The discourse on this topic has shown the need to quantify the mental effort of IS users and postulated the need for research into mental effort measures based on objective parameters such as behavioral signals, eye scanning movements, or physiological variables [7]. Also the discussions about metrics for human-robot interaction emphasised the need for research into a more objective mental effort measurement technique (e.g. “At this point in time, there is a need to identify non-intrusive measures of workload...” [18, p. 38]). NeuroIS researchers proposed determining a user’s mental effort based on objective psychophysiological measurements [5]–[7].

Since pupillary assessment is an unobtrusive nonreactive research method well established in psychology [19,20] and physiology [21,22] and recently also used in IS research [11,23]–[29], we will focus in what follows on the anatomical mechanism of the pupillary to show that mental mental effort directly leads to a pupillary dilation.

2.2. The mechanism of mental effort-based pupillary dilation

The pupillary diameter is determined by the tone of the iris dilator muscle (radial muscle of iris, radiating fibers) and the iris constrictor muscle (iris sphincter muscle, circular fibers) [30]. The pupil dilates when the dilator muscle is stimulated or the constrictor muscle is inhibited [20, p. 19]. The iris dilator muscle acts as an antagonist to the pupillary iris constrictor muscle. The innervation of the iris dilator and the iris constrictor muscles are controlled by the vegetative (involuntary) nervous system, which work largely below the level of consciousness. That is why the pupillary diameter responses evoked by specific stimuli occur spontaneously and are very difficult to manipulate over a longer period of time [20, p. 19]. The vegetative nervous system consists of the sympathetic nervous system and the parasympathetic nervous system.

The sympathetic nervous system is mainly involved in the fight-or-flight response [31]. In particular, the iris dilator muscle is stimulated by the sympathetic nervous system. The parasympathetic nervous system is mainly involved in the body’s preparation for rest and digestion issues. In more detail, the iris constrictor muscle is controlled by the parasympathetic nervous system that originates from the Edinger-Westphal nucleus [30]. Mental effort was found to inhibit the activity of the Edinger-Westphal nucleus which directly leads to a pupillary dilation [32].

2.3. Pupillary responses as a marker of mental effort in psychological research

Before the mental effort-based inhibition of the Edinger-Westphal nucleus was found, psychologists observed the relationship between mental effort and pupillary dilations. The initial work on task-evoked
pupillary responses stemmed from Hess and colleagues [33] as well as Kahneman, Beatty and colleagues [34]–[36]. The corresponding experiments were based on arithmetical, listening and talking tasks. Results coherently indicated a relationship between humans’ dedication to tasks and their pupillary diameters. In more detail, at this early stage of research most of the psychologists involved only assumed the link between cognitive workload and the pupils’ dilation. Further attentional processes had been identified, but could not be satisfactorily explained.

However, following the work of Kahneman and Beatty [34], Bradshaw [37,38], and Simpson [39] researchers began to realise that the pupil dilates long before task-evoked cognitive workload occurs and the pupillary response is more involved in individuals’ dedication/effort/attention processes [40]. The finding that the pupillary diameter response represents the mental effort and dedication of the individual was confirmed through the identification of the “preparedness effect”, which occurred before any demanding task.

Initial evidence that the pupil dilates long before cognitive workload occurs, which indicates the “preparedness effect”, came from Hakarem and Sutton [41]. They found that the pupillary dilates more when a report is required later compared to tasks where a report is not required. Hakarem and Sutton had already suspected that “this difference in dilation may reflect the different levels of vigilance required in the ‘report’ and ‘no report’ conditions” [41, p.485]. Simpson [39] found something similar, i.e. that a subsequent indication of task completion causes a higher pupillary dilation during a preceding cognitive task. This effect could not satisfactorily be explained and it was assumed that the individual is more excited in anticipation of later task evaluation and recognition. Simpson and Molloy [42] revealed that pupils of participants with audience anxiety are much larger compared to participants without audience anxiety. Based on these indications and their own results, Kahneman and Wright [43] noted that the “pupillary diameter ... is a measure of the intensity of mental effort” [43, p.188]. “Further, as in the other studies which showed a preparation effect on pupil size, this effect vanishes completely once actual work begins.” [43, p.189]. The so-called “preparedness effect” of the pupillary, as it was originally called by Kahneman and Wright [43], was also later found by Stanners and Headley [44].

Beatty [45] reviewed the existing empirical work concerning task-evoked pupillary responses and concluded that the pupillary dilations, which occur before and after task-execution, belong to certain attentional processes [45, pp.283]. He assumed that the pupillary dilates as a function of task-preparation and remains widened due to sustained attention. In a later experiment, Richer and Beatty [46] analysed pupillary responses on self-triggered finger flexion experiments. The pupillary response began about 1.5 seconds before the motoric finger action started – indicating a shift of attention to task-preparation. Richer and Beatty [47] revealed different reaction latencies of the pupillary diameter response peak, depending on task complexity, while the pupil diameter peak value itself was not effected. Since the pupillary diameter signal slowly increases just before task-presenting, results indicated that the signal reflects interest/dedication/attention shifts as a function of task-preparation. Qiuyan et al. [48] found that a task-evoked pupillary diameter response “in random stimulus sequences is sensitive to the surprise value of events”, which also indicates the task-preparation of the individual.

2.4. Pupillary responses as a marker of mental effort in IS research

In recent years, IS researchers also began to use the relationship between mental effort and the pupillary.

Rudimentary work (based on games, simple/trivial (arithmetic) tasks, non-evaluated frameworks, etc.) stems from [11,23]–[27]: Pomplun and Sunkara [23] (n = 10) used pupillary dilation as a mental effort indicator within a simple visual experiment asking users to find numbers in ascending order and to read them out loud. Longo [25], in a research in progress work, sketched out a very rudimentary framework for mental effort assessment using information technology. Cegarra and Chevalier [24] (n = 4) experimentally evaluated the mental effort of users solving a Sudoku puzzle by capturing pupil diameter data from eye-tracking. Xu et al. [26] experimentally studied pupillary responses indicating mental effort when performing arithmetic tasks given by a computer under luminance changes. Wang et al. [27] (n₁ = 13, n₂ = 12) investigated pupillary response as a mental effort measurement under the influence of different luminance levels and emotional arousal. The experimental setup contained simple arithmetic tasks and memorizing/reproducing tasks. Buettner [11] (n = 12) reported results from an experiment which determined the state of a user’s mental effort state based on the analysis of pupillary Hippus (continuous small fluctuations) using eye-tracking technology.

The little IS-related work on the task-evoked mental effort - pupillary diameter relationship using a more realistic experimental setup, stems from two groups – Iqbal and colleagues [49,50], and Buettner and col-
leagues [28,29]: Iqbal et al. [49] \((n = 12)\) and Bailey and Iqbal [50] \((n = 24)\) measured changes in mental effort during the execution of different types of tasks (route planning [49,50], document editing [49,50], E-mail classification [49]). Buettner [28] \((n = 5)\) investigated whether a higher level of artificial intelligence support leads to a lower user mental effort. Buettner et al. [29] \((n_1 = 12, n_2 = 125)\) reliably detected three different levels of mental effort in users.

In summary, it is noticeable that prior research on physiology-based measuring a users mental effort is very rudimentary (games, simple/trivial [arithmetic] tasks, non-evaluated frameworks, etc.). The few investigations using a more realistic experimental setup build on the work of Iqbal and colleagues [49,50] and Buettner and colleagues [28,29]. In addition, almost all experiments are based on a few participants \(4 \leq n \leq 32\). The sole large-scale experiment stems from Buettner et al. [29] \((n = 125)\). But the most pressing problem is that all prior investigations evaluate the users mental effort \textit{ex post} the experiment. There is no design-oriented work that has built a pupillary-based mechanism to reliably detect mental effort states in real-time – despite almost all researchers stressing in their discussions that their pupillary-based measurement approaches are suitable to predict it in real-time. Because there is no doubt that action-oriented NeuroIS needs such a mechanism, as many researchers have emphasised in the past (e.g. [6,51,52]), we are the first to build one. Note that our eye-tracking based approach is cost efficient, non-invasive and contact-free. EDA, HRV, EEG, fEMG are not contact-free measurements and fMRI, PET or MEG are very cost-expensive [7, pp. A1].

3. Methodology

3.1. Applying the NeuroIS guidelines

In order to clearly contribute to NeuroIS research and show strong methodological rigour we strictly followed the NeuroIS guidelines of vom Brocke et al. [53]. In particular, to assess prior research in the field of measuring mental effort as an important IS construct a comprehensive literature review was conducted (cf. [54]). To base our experimental design adequately on solid research in related fields of neuroscience [53] we showed the fundamental anatomic mechanism of the pupillary dilation controlled by the vegetative nervous system and the key role of the Edinger-Westphal nucleus that is inhibited by mental effort and directly leads to a pupillary dilation (see section 2.2). Further we showed in-depth experimental results from psychology on task-evoked pupillary diameter responses and existing related work in IS research (see sections 2.3, 2.4).

Our new methodology uses eye-tracking-based pupillometry as a well established approach in physiology [21,22] and psychology [19,20] “widening the ‘window’ of data collection” [19, p. 93]. With our method bio-data (i.e. pupil diameter) can be used to better understand mental effort as an IS construct (cf. guideline 4 of [53]). In comparison to other neuroscience tools eye-tracking-based pupillometry is the contact-free and efficient method of choice [20].

We applied the guidelines and standards from Duchowski [55] and the Eyegaze Edge™ manual.

3.2. Research model and hypothesizing

As described in section 2.3, in engaged individuals the pupillary dilates before a simple task is presented (cf. [34,37]–[40]). In more realistic tasks individuals not only have to solve one simple task such as mental arithmetic but a sequence of sub-tasks (e.g. orientation, information search, reading, recall of information, writing). This sequence of sub-tasks leads to a recurring demand of mental activity resulting in a variability of the pupil diameter. That is why in realistic task settings the measurement of interest is not only the pupil diameter itself but also its variability. Because of this reason, prior research using realistic task-environments applied pupil diameter variability measures such as the percent change [49,50] and standard deviation [29]. The (early and sustained) pupil diameter variability (PDV) as the recurring pupillary diameter response represents the individuals’ dedication/effort in a successful task fulfillment and may be a suitable predictor of user performance. Against this background, we hypothesize that:

\[ H_1: \quad \text{The pupil diameter variability is a predictor of user performance.} \]

3.3. Laboratory setting and measurement

To capture the pupillary diameter as the measure of interest in this research, eye-tracking was performed using the binocular double Eyegaze Edge™ System eye-tracker paired with a 19” LCD monitor (86 dpi) set at a resolution of 1280x1024, whereby the eye-tracker samples the pupil diameter at a rate of 60Hz for each eye separately.

3.4. Description of the test procedure

In order to evaluate our method on a complex widely-used information system with a lot of dynamic
web elements, interaction functions and advertising banners, we chose the online social network application LinkedIn.com and performed a large scale experiment [56]. The participants in our experiment had to solve three tasks with a varied level of difficulty (low/mid/high) inducing different mental effort levels of the users:

1) “Ask your contact [given first name, surname] for a letter of recommendation.” (eight steps between login and task fulfillment → high demand level)
2) “Apply to Oracle for a new job of your choice.” (four steps between login and task fulfillment → low demand level)
3) “Identify and name the Global Head of Recruiting at BMW.” (six steps between login and task fulfillment → mid demand level)

These three tasks had to be fulfilled in the same test order. In order to counterbalance our design, we used the high demand task first, followed by the low demand task, followed by the mid demand task.

Prior to all data collection, each test participant is welcomed by the experimenter (supervisor of the experiment). After that, the participant has to fill out a consent form and also a questionnaire with demographics (stage 1). In stage 2, the supervisor turns the task-sheet, including a short note about each task that has to be fulfilled over to the participant and reads the tasks out aloud. In what follows, the participant has time to read each task again and to ask questions. In stage 3, we take the necessary precautions for the experiment, for which we make use of the eye-tracker. Hence, the eye-tracker is calibrated. In stage 4, the experiment starts with the first task the participant has to accomplish.

### 3.5. Sampling strategy

We recruited 129 participants from a pool of extra-occupational MBA and bachelor students. All of them had professional working experience, making them realistically suited to application situations. To ensure that all participants understood the scenario and the LinkedIn.com system, they were given introductions to the system and the computer interface.

### 3.6. Data cleansing

The eye-tracking system is able to track participants’ eye movements and pupillary responses when the participants are looking in the direction of the monitor. The more complex and realistic a task is, the more participants tend to look away from the monitor, e.g. at the keyboard, and the eye-tracking system is no longer able to record every movement of the eye. In real conditions, participants look in the direction of the monitor for about 70% of the time. In addition, some naturally determined artifacts occur, e.g. by eye blinks (cf. [57]). When the eye-tracker loses the participants’ eyes it marks this data time-accordingly (“invalid”).

We use this eye-tracking signal without any further data cleansing approach in order to show the robustness of our prediction method.

### 3.7. Random forest method

In this study, the Random Forest (RF) method was used to predict user performance using pupil diameter data. RF is a machine learning classifier which is based on an ensemble of unpruned decision trees [58]. The classification decision is based on a majority vote principle based on all trees of the RF. The conceptual idea underlying a decision tree is to recursively identify a predictor that allows the sample to be split in two subparts that are as homogeneous as possible with regard to the classification task at hand. For binary predictors (yes/no) the split point of the variable is self-evident; for polytomous or continuous predictors the algorithms identify the most selective split point for the dependent variable using e.g. entropy as a measure. In this way, a tree-like structure is built. The procedure is repeated until a stop signal is reached – e.g. all cases are classified, or the algorithm cannot improve the accuracy of the classification anymore [58]. Such types of algorithms are called recursive partitioning because the sample is subdivided (i.e. partitioned) into smaller parcels in a reiterated manner.

RF has been successfully applied to a number of different neuro- and bio-science related research problems such as brain imaging [59], gene expression [60], and biomarker identification [61] and in recent times to IS-problems [62]. In particular, RFs are especially useful in, but not limited to, “small n, large p” problems, where the number of predictor variables p is larger than the number of cases n. Even with sufficiently large samples RF can be a valuable tool, as they allow the delineation of statistical properties such as nonlinear trends, high-degree interaction, and correlated predictors. Additionally, assumptions that are needed for classical multivariate analyses such as homoscedasticity (homogeneity of variance), linear associations between variables, or metric variable levels are not necessary [58].

1. Buettner et al. [29] showed on the same data set that the standard deviation of the task-evoked pupillary diameter responses significantly correlated with the users mental effort level and with user performance ($p < 0.01$).
4. Results

On average we captured 63,600 pupillary data points from eye-tracking within the mean duration of 13 minutes of the experiment for each of the participants. The captured data comprises of time-stamped pupil diameter values for both left and right eyes.

4.1. Sample characteristics

125 of the 129 participants (97 %) could be appropriately calibrated with the eye-tracking system. The other four test participants, who could not be successfully calibrated due to problematic eyeglasses or visual defects, were deleted from the dataset. The remaining participants are aged from 21 to 61 years (M=29.6, S.D.=7.2). 56 persons are female, 69 male.

4.2. Pupil diameter time series results

The time series for the pupillary diameter values are presented in figure 1 for the left eye and in figure 2 for the right eye.

The results shown in both figures 1 (left eye) and 2 (right eye) clearly indicate that the outperformers have a higher pupillary dilation compared to underperformers, which indicates a higher mental effort by the outperformers. In addition the pupillary dilation of the outperformers was sustained better over time – also indicating a higher mental effort level.

4.3. Relationship between pupillary diameter variability and task difficulty

In order to evaluate the hypothesis $H_1$ we calculated the standard deviation (S.D.) of the median pupil diameter (the measure of pupillary diameter variability, PDV) as the recurring pupillary diameter response for each task (cf. [29]).

As shown in table 1 we found a strong relationship between the pupillary diameter variability and the difficulty of a task indicating that a higher task difficulty demands more variability of the pupils.

<table>
<thead>
<tr>
<th>Task difficulty</th>
<th>Success rate (%)</th>
<th>PDV left eye [mm]</th>
<th>PDV right eye [mm]</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>65% (81 of 125)</td>
<td>0.08296</td>
<td>0.07891</td>
</tr>
<tr>
<td>med</td>
<td>34% (42 of 125)</td>
<td>0.09025</td>
<td>0.08880</td>
</tr>
<tr>
<td>high</td>
<td>22% (28 of 125)</td>
<td>0.09559</td>
<td>0.09949</td>
</tr>
</tbody>
</table>

The differences of the pupillary diameter variabilities between the low and the medium mental demand level as well as between the medium and the high mental demand level were all significant ($p_{left} < 0.01$, $p_{right} < 0.001$). This result supports hypothesis $H_1$.

4.4. Relationship between pupillary diameter variability and performance

In order to assess the relationship between PDV and user performance, PDV was calculated as the standard deviation (S.D.) of the median pupil diameter for each of the 125 participants.

As a further result we found a clear relationship between the performance score and the pupillary diameter variability (table 2). Higher performance scores corresponds with higher PDVs – supporting
our hypothesis $H_1$. All differences between the absolute underperformers (0 score) and other performance groups (1, 2, or 3 score) were significant ($p_{left} < 0.1, p_{right} < 0.05$). In addition, the difference between the 1 score group and the 3 score group was significant for the right eye ($p_{right} < 0.1$).

Table 2: Relationship between pupillary diameter variability (PDV) and performance score

<table>
<thead>
<tr>
<th>Performance Score</th>
<th>PDV [mm] left eye</th>
<th>PDV [mm] right eye</th>
</tr>
</thead>
<tbody>
<tr>
<td>0/3 - no correct task</td>
<td>0.21549</td>
<td>0.22182</td>
</tr>
<tr>
<td>1/3 - one correct task</td>
<td>0.28828</td>
<td>0.30699</td>
</tr>
<tr>
<td>2/3 - two correct tasks</td>
<td>0.29199</td>
<td>0.31798</td>
</tr>
<tr>
<td>3/3 - all tasks correct</td>
<td>0.30284</td>
<td>0.36028</td>
</tr>
</tbody>
</table>

4.5. Random forest analysis

As the participants needed different amounts of time to complete the three tasks, the number of measurement points differed between participants. As the employed algorithm relied on data vectors of the same length, and in order to reduce the data to a manageable size, we compressed the measurement points to 1,000 measurement points per eye per participant (mean compression factor: 95.67; S.D.: 49.09). Precisely, for each participant, one compressed set of measurement points was based on the median of the adjacent uncompressed measurement points.

Next, we performed a RF for each measurement point using 2,000 trees per RF using standard settings. We used the R environment 3.1.0 for all analyses [63]; and we used the package randomForest for the RF analyses [64]. Our R source code can be downloaded from dx.doi.org/10.6084/m9.figshare.1050342.

One flaw of decision trees is their instability with regard to the input data. As a consequence, relatively small changes in the input data may have a great effect on the classification results. This undesirable effect may be avoided by drawing random samples both of cases and of features. For example, in a typical RF more than 1,000 trees may be “grown” thereby diminishing the effect of strong influential individual cases resulting in more stable results. The effect of drawing a random subset of variables that are considered as entry parameters for an individual tree (the excluded variables cannot enter the respective tree) allows “weaker” information to enter the model - potentially important interactions that would otherwise be concealed by one or more powerful variable(s) and would remain unidentified by the analysis. It is therefore important that an analytical framework – RF or otherwise – should not be tested on the basis of its own construction sample because the danger of overfitting is present. For this reason, the predictive accuracy of a decision tree in the RF method is estimated by the part of the sample which was not used for building the RF framework. This excluded part of the sample is called the “out of bag” (OOB) sample and its advantage is that the risk of overfitting is avoided. Thus, the predictive quality of RF is assessed by means of a cross validation sample-approach. It is noteworthy, that for each tree, cases are selected by drawing with replacement, thereby including approximately $1 - e^{-1} \approx 2/3$ of the cases at each tree. Here, 63% of the data was used for the training sample, 37% for the validation sample.

4.6. Predictive model quality

In our RF model we were able to further distinguish between the outperformers more granularly by predicting class membership. As only three participants scored 0 points, we combined score 0 and score 1. This decision was based on the fact that the RF is unable to learn using particularly small sample sizes. Our RF algorithm for performance predicting discriminated well between the three classes as the out-of-bag (OOB) error values were substantially lower compared to the random line of $2/3$ (OOB$_{left}$ = 0.54 $\ll$ 0.67, OOB$_{right}$ = 0.58 $\ll$ 0.67).

Figure 3: Mean OOB prediction error for Random Forests for measurement points 1-1.000 for left eye (upper panel) and right eye (lower panel).

As shown in figure 3 the OOB error values were already very low within the first quarter of the normalized trial time (1–250 thousandths). That is impressive because at this point of the trial time no task was completed but a good predictive performance with our RF algorithm is possible. The prediction accuracies based on the left as well as the right pupillary response were substantially higher compared to the random line of $1/3$ (ACC$_{left}$ = 0.46 $\gg$ 0.33, ACC$_{right}$ = 0.44 $\gg$ 0.33). The RF algorithm reached a good overall correct prediction rate (ACC$_{overall}$ = 0.45 $\gg$ 0.33).

These results support our hypothesis $H_1$. 

539
5. Discussion

As anticipated, our experimental data revealed clear differences within the pupil diameter time series between outperformers and underperformers. This indicates that pupillary diameter data may a possible predictor for user performance. In addition we found a strong relationship between the pupillary diameter variability and the difficulty of a task – supporting \( H_1 \). In fact, a higher task difficulty demands more variability of the pupils. We further showed that there is a clear relationship between the user performance and the pupillary diameter variability – also supporting \( H_1 \). Our results from algorithm-design by using the Random Forest Method showed a good ability to predict the user performance – also supporting \( H_1 \).

Our proposed method is very robust and was not affected either by luminance differences within the regular band of in-room lighting conditions or by age differences. This is impressive because the pupil size is primarily influenced by luminance caused by lighting conditions [22], but also by an individual’s age.

In addition, our results are very promising because we evaluated our method within a realistic large scale experiment on the online social network application LinkedIn.com as a complex widely-used information system with a lot of dynamic web elements, interaction functions and advertising banners.

6. Conclusion

We contribute to IS research by being the first to propose a method of ex ante predicting user performance based on a non-invasive contact-free cost-efficient real-time measurement of mental effort. Our new methodology uses eye-tracking-based pupillometry capturing pupillary diameter data and calculates – based on a Random Forest algorithm – user performance expectations. Based on a realistic large-scale experimental evaluation of our methodology we showed promising results using the pupillary diameter (variability) in order to predict user performance.

With our new methodology we are able to measure mental effort in real-time and do so more objectively which are two big advantages compared to a users’ subjective self-rating. Based on this measurement our methodology also predicts user performance in real-time. Our results are very promising and allow the dynamic adaption of IT to a user’s mental effort and the expected user performance (e.g. regulating the degree of information complexity, the font size or by presenting help texts) – in real time – and thereby address one of the most challenging problems in IS research. User performance belongs to the most important IS constructs [1] and mental effort is in all probability closely related to important IS constructs such as perceived ease of use (cf. [3,4,6,8]).

Our methodology can be applied cost-efficiently to both research and practical environments without disturbing the participant/user.

6.1. Limitations

Our main limitation is rooted in the use of RF as the sole machine learning (ML) approach. We decided to use RF for a variety of reasons: As we do not use a complicated data cleansing approach in order to guarantee the usability of our methodology independent of the chosen eye-tracking system we recognized outliers in pupillary data. But RF is not very sensitive to outliers in training data [62]. In addition, overfitting is not a problem in RF and the RF-algorithm learns very fast [58] – two important preconditions for real-time predictions. Finally, RF regularly leads to greater accuracy than simple/mixed effect regression models or other classifiers [62]. However, an evaluative comparison of other ML approaches is still a possibility.

In addition, we tried to counter-balance our research design by using the high demand task first, followed by the low demand task, before ending with the mid demand task. But a randomized task sequence instead of a fixed one (high/low/mid) remains to be tested.

6.2. Future research

Our research opens a lot of future empirical and design-oriented research opportunities. With our cost-efficient and physiology-based methodology it is possible to empirically investigate the relationships between mental effort and other important IS constructs such as perceived ease of use in a much more sophisticated manner (in accordance to [6]–[8]). In addition new research opportunities open due to our methodology concerning a dynamic adaption of IT to users’ mental effort. Such NeuroIS-driven research substantially enhances our understanding of IS constructs and ensures the technological impact of IS research results.

However, as described in the last section, future work should evaluate other ML-algorithms such as support vector machines, Bayesian networks, and artificial neural networks concerning its predictive power of user performance based on pupillary diameter variability. In addition, future work should investigate alternative means for measuring the variance of the pupillary diameter (e.g. various statistical variance measures).
The relationship found between PDV and user performance (see table 2) needs further investigations since the user performance-based increase of PDV was not always significant. Future research should deepen the analysis concerning a non-linear relationship (e.g. saturation effects of PDVs).

Acknowledgments

We would like to thank the reviewers, who provided very helpful comments for refining the paper.

References


