Physiological Responses to Well-Designed and Poorly Designed Interfaces

Robert Ward, Phil Marsden, Bernadette Cahill and Clive Johnson
School of Computing and Mathematics
University of Huddersfield
Huddersfield, HD1 3DH, UK
+44 1484 472000
[r.d.ward | p.h.marsden | b.cahill | c.a.johnson] @hud.ac.uk

ABSTRACT
Physiological indicators of arousal have long been known to be sensitive to mental events such as positive and negative emotion, changes in attention and changes in workload. It has therefore been suggested that human physiology might be of use in the evaluation of software usability. To this, there are two main approaches or paradigms: 1) periodic measures, i.e. comparisons of average readings across periods of time as measures of different arousal levels under different circumstances, and 2) local measures, i.e. the detection of sudden changes in arousal that may have been brought about by particular events.

This paper reports the results of an experimental investigation in which participants carried out a web-based task using either a well-designed interface or a poorly-designed interface to exactly the same content. Participants using the poor interface showed higher levels of arousal throughout the task (paradigm 1) which could in part be attributed to specific discrete HCI events (paradigm 2). The outcomes support the contention that psychophysiology can be informative about software usability, and illustrate a viable approach to investigation in this area. However, further experiments involving non-contrived tasks are now called for.

Keywords
Affective computing, psychophysiology, usability testing, content evaluation.

INTRODUCTION
Physiological changes in skin conductivity, heart activity, blood pressure, respiration, voice quality, eye movements and electrical activities in muscle and brain, have long been known to occur in response to thought processes and mental events [1]. These phenomena, which reflect arousal, are both involuntary and surprisingly sensitive. It has therefore periodically been proposed that they may be of use in helping to identify HCI events of significance to users. Physiological changes would seem likely to accompany different kinds of emotional response, for example negative emotion such as frustration caused by software problems, positive emotion such as delight on completing a particular task, and shifts in attention in response to particular content and moments of high workload.

A number of studies have provided support for the application of this general idea in computer-related situations. These essentially fall into two main approaches or paradigms. Firstly, there have been studies that make comparisons between repeated psychophysiological measurements averaged across periods of time with the aim of demonstrating different levels of arousal in one situation as compared with another. For example, cardiovascular measures have been found to indicate reductions in the stress levels of ambulance control system operators at times of high workload when a computer-based system replaced a manual, paper-based one [7]. Similarly, skin conductivity and cardiovascular measures have been found to indicate increased stress levels in viewers of video following a change from a high to a low frame rate, even though many participants were unaware that there had been a decrease in video quality [8]. Secondly, there have been studies that make use of relatively sudden psychophysiological changes, referred to as orienting responses, occurring either immediately after a known event, implying increased arousal as a result of that event, or occurring freely, suggesting that something has happened to cause a sudden change in arousal. For example, sudden changes have been reported in the muscle electrical activity of computer game players, especially in situations where software fails to react correctly to its controls [4], and the present authors have previously reported that psychophysiological responses reliably occur in response to certain kinds of discrete HCI events [6].

Unfortunately, both paradigms involve difficulties in methodology, data analysis and interpretation of results, mainly that (i) physiological measurements typically show differences between measures, and between and within individuals, (ii) defining and identifying orienting responses

in terms of the properties of physiological signals involves somewhat arbitrary assumptions such as "a change of 5% within 2 seconds", (iii) even when a physiological response is clear it could indicate any one of several different emotional reactions depending on the situation in which it occurs. Thus, although there is empirical support for the idea of employing psychophysiological measurement to identify significant HCI events, difficulties of methodology, data analysis and interpretation call into question its viability. Further investigation is therefore needed.

This paper reports an experimental investigation of physiological responses to different interface designs. Physiological measurements were taken from participants engaged in a task using two different versions of a web site, a well-designed version and a poorly-designed version. The aims of the investigation were (i) to trial an experimental procedure for investigating psychophysiological aspects of software use, (ii) to investigate physiological changes in response to software use over time (paradigm 1), to (iii) to investigate physiological changes in response to discrete interface events (paradigm 2).

METHOD

An experimental web-based task was devised, based upon an unpublished digitised directory of organisations and residents of a Yorkshire town and its surrounding villages for the year 1939 [5]. This directory was HTML-based and delivered by Microsoft Internet Explorer. It consisted of a front index page providing links to scanned images of the 340 pages of the original print version of the directory. Two interfaces to the directory were created, a "well-designed" interface and a "poorly-designed" interface, both accessing exactly the same information. The well-designed interface followed as far as possible principles of good web and information design [2, 3]. The poor interface broke many of these principles in ways all of which have been observed in genuine web sites. Its features included:

- An index page that made excessive use of pull-down lists, which obscured the overall structure of the information in the directory, making links difficult to find and use.
- Impoverished navigational cues and functions so that navigation involved additional scrolling and other mouse operations.
- Gratuitous animation and periodic advertisements which either (a) produced pop-up windows that had to be moved or closed in order to proceed, or (b) caused screen content suddenly to change position, forcing users to make unanticipated adaptations in their mouse targetting movements.

For the investigation, 20 participants aged 18 to 48, drawn from the general University population, were alternately assigned to either the "well designed" or "poorly designed" version of the directory. Participants were required to work through a series of questions that required them to find information in the directory (e.g. "How many people named Young lived at Rawcliffe?", "Who was the clerk of Airmyn Parish Council?"). Skin conductivity and blood volume pulse were measured over a 15 minute period consisting of an initial 5 minute "settling in" period, followed by 10 minutes on the question answering task itself. It was hypothesised that the "poorly designed" version would be more difficult to use, and that this would be reflected in the physiological data.

Signal data was collected using Lafayette's DataLab 2000, with analysis carried out with National Instruments BioBench and Microsoft Excel. Skin conductivity (GSR) and Blood Volume Pulse (BVP) were measured through electrodes and sensors attached to the fingers of participants' non-dominant hand (leaving the dominant hand free to carry out experimental tasks). The BVP sensor also detects pulse, and therefore provides a measure of heart rate (HR). Two Windows PCs were used, one for data collection and one to present the experimental task.

RESULTS

Skin conductivity, peripheral blood volume and pulse rate over the first minute of the question task was taken as the baseline reading for each participant. Each participant therefore acted as their own control, avoiding methodological difficulties caused by individual differences and occasion specificity. Changes against this baseline were plotted over the subsequent 9 minutes. This in effect provides a measure of recovery from the initially high levels of arousal that occur at the start of a task.

All three measures showed mean differences in the direction predicted by the hypothesis. Users of the well designed interface showed mean decreases in skin conductivity of 2.30% over the period of the task, compared with mean increases of 2.23% in users of the poor interface. Users of the well-designed interface also showed a greater mean increase in finger blood volume of 7.0% compared with 3.2% in the poor group, and a mean decrease in pulse rate of 2.7 beats per minute compared with 0.3 beats per minute in users of the poor interface.

Figure 1 plots the data as group percentage changes against baseline over each minute of the task. This indicates that the skin conductivity of users of the well designed interface began to decrease after the first minute, whereas for users of the poor interface it increased over each of the following four minutes. Only after the seventh minute did skin conductivity show marked decline for the poor interface group. Finger blood volume shows a similar trend, but does not differentiate so well between the conditions. Pulse rate in the well-designed interface group dropped quite quickly after the start of the task and then remained low, whereas in the poor interface group it continued at around the initial level throughout the task.

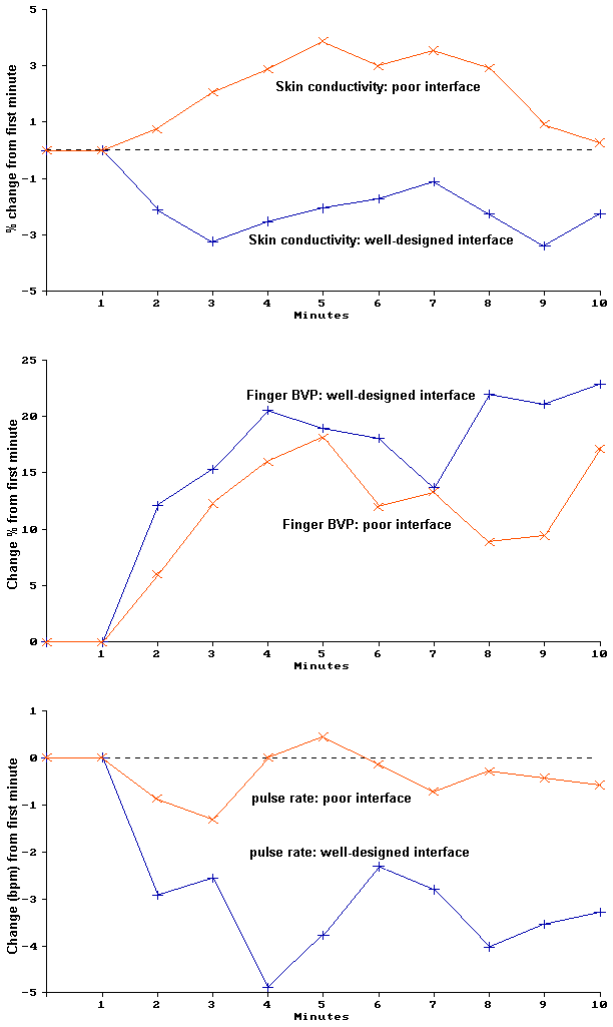

FIGURE 1: group changes in skin conductivity, finger blood volume and pulse rate over each minute of the experimental task.

The group means summarise considerable individual differences between participants. There are large variances in group data, and the groups contain individuals with large values pulling in opposite directions. Using the strictest statistical criteria the group differences are not therefore statistically significant. This is discussed further in the conclusions.

The data was also examined for indications of changes in skin conductivity in response to discrete events. Skin conductivity was found to show a significant difference immediately before and after the appearance pop-up advertisements in the poor interface. Pop up advertisements appeared 93 times across all participants using the poor interface. During the 15 seconds prior to appearance of advertisements, participants' skin conductivity decreased by a mean of 0.0308 microSiemens, compared with an increase of 0.026 microSiemens in the following 15 seconds, an overall mean difference of 0.057 microSiemens, approximately 0.9% of the conductivity levels prevailing at the point that advertisements appeared. This change in skin conductivity was found to have occurred in all participants, and was statistically significant across the 93 events ($t=2.255$, $p<=0.05$).

CONCLUSIONS

The results of this investigation are promising. The group summary data indicates that measures of skin conductivity, blood volume and pulse rate, averaged across periods of time, are able to distinguish differences in arousal levels in different computer-based situations, and can therefore provide an indication of software usability. The data also provides further evidence that statistically significant differences in reactions to discrete HCI events can be detected. The general approach and procedure for usability testing adopted in the investigation appears viable and applicable to different software in different HCI situations.

Although the results are promising, they should be regarded as indicative and taken with caution because of the large group variances. There would appear to be a number of possible contributory factors to these variances. Some individuals simply seemed to show greater reactivity than others. Participants also appeared to differ in their ability to handle the usability problems posed by the "poorly designed" version. It seemed that the more experienced web users tended to proceed slowly and methodologically, without becoming too irritated by the difficulties. There were also other uncontrolled factors. Some participants attempted to engage the experimenter in conversation during the experimental procedure. Users of the well designed interface answered around twice as many questions as those using the "poor" version, and some of the later questions were relatively more difficult and required more mouse clicks. Thus task and workload effects may have differed between the groups. These factors reflect that the nature of the investigation was of pilot exploration rather than rigorous experimentation. Figure 1 is strongly suggestive of differences between the groups, and it is hypothesised that the results of a replication under more tightly controlled conditions currently being carried out will show statistically significant differences.

Interpretation of results was flagged as problematic in the introduction. One appealing interpretation of the psychophysiological readings here is that they indicate relaxation into a task. Initially, on starting the task, participants are probably anxious about what the experimental situation requires of them. Users of the well designed interface appear soon to recognise that the task is a simple matter of finding the right page and reporting the answer. They relax, and as they do so their eccrine secretions decrease as indicated by lower skin conductivity, their peripheral blood vessels dilate as indicated by higher finger blood volume, and their pulse rate decreases. In contrast, users of the poor interface do not find the task so straightforward. They encounter difficulties and inconsistencies in the software that prevent rapid

understanding of the task requirements. One of these difficulties is having to deal with the unpredictable appearance of pop-up advertisements, and the physiological data provides evidence for this. Participants using the poor interface are therefore uncertain for a much longer period about how to find the required information, and their arousal remains high, suggesting continued anxiety.

The experimental task used in the investigation presented a contrived situation containing elements deliberately designed to cause usability problems. The approach needs now to be applied in non-contrived, more realistic situations where it is not known beforehand whether usability problems exist or when they occur. Psychophysiology would be a long drawn out way to tell us the obvious, but if it proves able to identify important HCI elements of which users are not conscious, or which they tend to forget when providing standard usability feedback, it could be an informative and valuable tool.

ACKNOWLEDGEMENTS

This work is supported by EPSRC project grant GR/N00586 and the University of Huddersfield.

REFERENCES

1. Andreassi, J.L. Psychophysiology : Human Behaviour and Physiological Response. Oxford University Press, 1980.
2. Hartley, J. Designing Instructional Text (3rd edition). Kogan Page, 1994.
3. Nielsen, J. Alert Box Columns 1995-2001. http://www.useit.com/alertbox/ (accessed 27th June 2001)
4. Picard, R.W. Affective Computing. The MIT Press. Cambridge, Massachusetts, 1997, p164.
5. Ward, R.D. Unpub. digitisation of “Goole 1939. The Almanack, Directory and District Year Book for the Year 1939”. Original published by The Goole Times Co. Ltd., Goole, 1939.
6. Ward R.D., Marsden P.H., Cahill, B. and Johnson. C.A. Using skin conductivity to detect emotionally significant events in human-computer interaction. Proceedings of IHM-HCI 2001, the joint 13th Annual Conference of the Association Francophone d’Interaction Homme-Machine (AFIHM) and 15th Annual Conference of the Human-Computer Interaction Group of the British Computer Society, II, 25-28.
7. Wastell, D.G. and Newman, M. Stress, control and computer system design: a psychophysiological field study. Behaviour and Information Technology Vol. 15, No. 3, 1996, pp. 183-192.
8. Wilson, G. and Sasse, M.A. Do Users Always Know What’s Good For Them ? Utilising Physiological Responses to Assess Media Quality. In S. McDonald, Y. Waern & G. Cockton [Eds.] : People and Computers XIV - Usability or Else! Proceedings of HCI 2000, pp. 327-339. Springer.