Emotional Investment in Naturalistic Data Collection

Ian Davies and Peter Robinson

Computer Laboratory, University of Cambridge, UK

Abstract. We present results from two experiments intended to allow naturalistic data collection of the physiological effects of cognitive load. Considering the example of command and control environments, we identify shortcomings of previous studies which use either laboratory-based scenarios, lacking realism, or real-world scenarios, lacking repeatability. We identify the hybrid approach of remote-control which allows experimental subjects to remain in a laboratory setting, performing a real-world task in a completely controlled environment. We show that emotional investment is vital for evoking natural responses and that physiological indications of cognitive load manifest themselves more readily in our hybrid experimental setup. Finally, we present a set of experimental design recommendations for naturalistic data collection.

1 Introduction

As control environments become increasingly complex, understanding the level and effects of cognitive load is becoming ever more important. Drivers are expected to operate their satellite navigation, telephone and in-car entertainment systems without compromising the safety of their driving. Aeroplane cockpits are providing more information to pilots flying in more crowded skies. Air traffic controllers must manage the increasing traffic safely and design the next generation control systems to help their operators avoid cognitive overload. It is essential that we build up an understanding of the effects of cognitive load on human operators, otherwise we rely on guesswork when designing new, usable systems.

This need for understanding has not gone unnoticed. Twenty years ago people were already investigating the effects of stress on physiological responses such as skin conductance and heart rate [2]. These studies were most commonly conducted in laboratory settings with artificial scenarios designed specifically to raise cognitive load in a particular way, such as overloading working memory [3]. Later, as technology for data collection progressed, studies began to focus on real-world scenarios. Healey, Picard and others have conducted in-car studies showing that varying driving conditions can cause different physiological responses [6,10] and that classifiers can be built to identify the conditions based on responses [5]. Lisetti and Nasoz have primarily conducted simulator-based experiments investigating the link between affective states and physiological indicators [8].

S. D´Mello et al. (Eds.): ACII 2011, Part I, LNCS 6974, pp. 467–476, 2011.

These studies were extremely important, initially to show that there is indeed a link between cognitive load and physiological responses, and latterly to verify that these effects can be observed in real-world tasks. Unfortunately the on-road experiments make quantification of the stimuli and effects extremely difficult because of the limited control over experimental conditions, particularly traffic quantity and behaviour. There is also the ethical difficulty of deliberately inducing stress in a driver surrounded by real traffic. Simulator-based studies give researchers the necessary control over the environment, but, as we shall see, may not be suitable for naturalistic data collection due to a lack of emotional investment on the part of the subjects. The issue of naturalistic data collection has come to the fore recently as people start to consider the use of real-world versus acted data in affective computing systems [11].

Here we present two experiments which explore this issue. The first is a simulator-based study of driver stress, where we found that for many subjects physiological responses are not a good indicator of cognitive load. We then hypothesised that participants often did not care about their performance when the scenario was completely artificial and we designed an experiment that provided a real-world control task but maintained our control over the environment.

2 Simulated Car Driving

The objective of the first experiment was to use a driving simulator to replicate results observed by others in real-world scenarios [5,6]. In particular, we hoped to confirm skin conductance and heart rate as good indicators of cognitive load and to investigate correlations with other, less invasive measures such as eye movements.

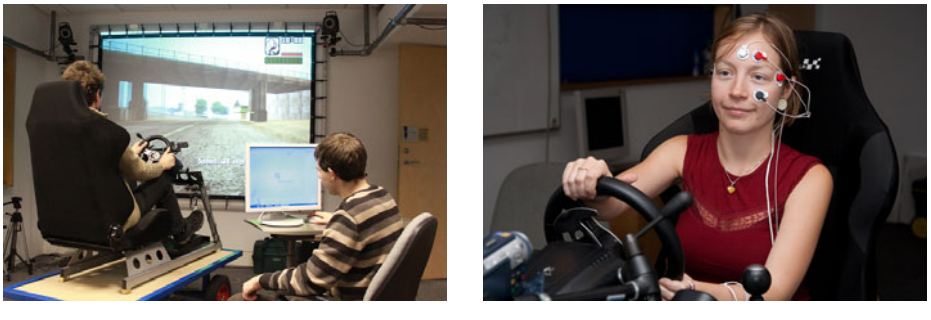


Fig. 1. Participants were asked to drive through several scenarios in our fixed-base driving simulator while wearing EOG, GSR and BVP physiological sensors

2.1 Experimental Setup

Figure 1 shows our fixed-base driving simulator. We use a seat, steering wheel and pedals to give a realistic cab-like environment for our participants, and a

projection screen which largely fills the visual field of the driver (around 60° field-of-view). A single PC runs the simulation software which is a slightly modified version of Rockstar Games' "Grand Theft Auto: San Andreas". Modifications allow remote monitoring of controls, speed and road position, as well as control of traffic and weather conditions. The game includes a 6 km x 6 km map with over 500 km of roads, including motorways, city streets and country lanes, which allows long and varied scenarios to be designed. A second PC runs the remote monitoring and control software for the simulator and also controls navigation instructions and secondary tasks. A third and final PC records physiological data gathered from the driver through our NeXus-4 Wireless Physiological Monitoring system. All the computers are synchronised appropriately for logging purposes.

2.2 Subjects

Subjects were recruited through advertisements in the local community. Fifteen subjects participated (nine female, six male), providing twelve sets of complete results. Participants were aged between 20 and 60 years, with about two-thirds below 30 years. All held full European driving licence, and those that required glasses or contact lenses for driving wore them. None had participated in previous studies in our laboratory.

2.3 Procedure

On arrival, participants filled in a pre-study questionnaire for the collection of demographic data. Electrodes were attached for measurement of skin conductance (GSR), blood volume pulse (BVP) and eye movement (EOG) and then they were given time to familiarise themselves with the simulator. They practised for as long as they wanted (typically around 15 minutes) until they were comfortable with the controls and the environment.

The main experiment consisted of six 5-minute scenarios of varying difficulty. Each scenario had a different route, but all routes started on city streets leading to a short section of motorway and finally into country lanes. In all conditions navigation instructions were given verbally by the supervisor. The secondary task consisted of verbal arithmetic questions such as "23 plus 7" in the "simple maths" conditions and "19 minus 76" in the "complex maths" conditions. There were two different weather/traffic conditions (clear weather with no traffic and stormy weather with heavy traffic) and three different secondary task difficulties (navigation only, navigation with simple arithmetic and navigation with complex arithmetic). This gave a total of six scenarios which were randomly ordered for each subject. Between each scenario, participants filled in a questionnaire with questions based on the NASA-TLX self-reporting scheme for workload measurement [4]. They were then given several minutes to relax and drive through open countryside.

2.4 Discussion

The first data analysis step for this experiment was to validate the relative difficulty of the scenarios. Participants' self-report ratings of the scenarios in

several categories were ranked, and Friedman tests followed by Wilcoxon Signed Ranks tests showed that navigation plus complex arithmetic was ranked as being significantly more demanding ($p < 0.005$), hurried ($p < 0.05$), frustrating ($p < 0.005$) and stressful ($p < 0.01$). We now compare the results of the "easiest" scenario (clear weather, clear traffic, no arithmetic) to the results of the "hardest" scenario (stormy weather, busy traffic, complex arithmetic).

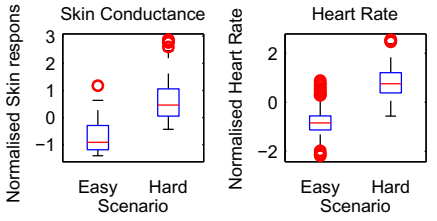
The results of physiological data analysis were less clear-cut. Some subjects showed strong skin conductance effects correlated with the scenario difficulty, but many bore no relationship. The same was true for heart-rate, blink-rate and heart-rate variability. Figure 2 (a) shows results from a person who showed indications of stress as expected, while Figure 2 (b) is typical of most of our subjects who showed no such response. Figure 2 (c) shows aggregated results from all participants. Note the lack of any significant effect. See Section 3.4 for a detailed description of the physiological data processing procedure. Note that data from several participants was discarded completely following hardware failure. Only complete, valid data sets were analysed.

We consider three possible causes for the disappointing results. Firstly, the possibility that having a supervisor in the room giving navigation instructions and asking arithmetic questions introduced a confounding effect due to social interaction. Secondly, the possibility that the five-minute scenarios were simply too short for the effects of high cognitive workload to manifest themselves. A brief follow-up study where the navigation instructions and secondary task were entirely automated using a speech synthesiser and the scenarios were extended to 20 minutes did not yield clearer results, suggesting that neither of these possibilities was a major source of problems.

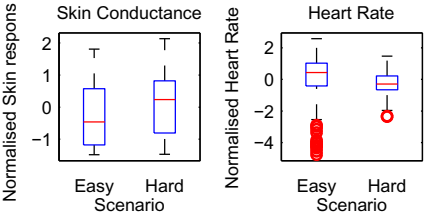
The third possible cause we consider is the potential for subjects to lack "emotional investment" in the task when using a simulator. In particular, many subjects would laugh when they crashed, suggesting that they did not really care about their driving performance to the same extent as they would have in on-road studies such as those conducted by Healey and Picard [5,6]. We suspected that this lack of emotional investment meant that although subjects may have been working hard, many did not exhibit the classic indications of stress.

3 Remotely-Controlled Flying

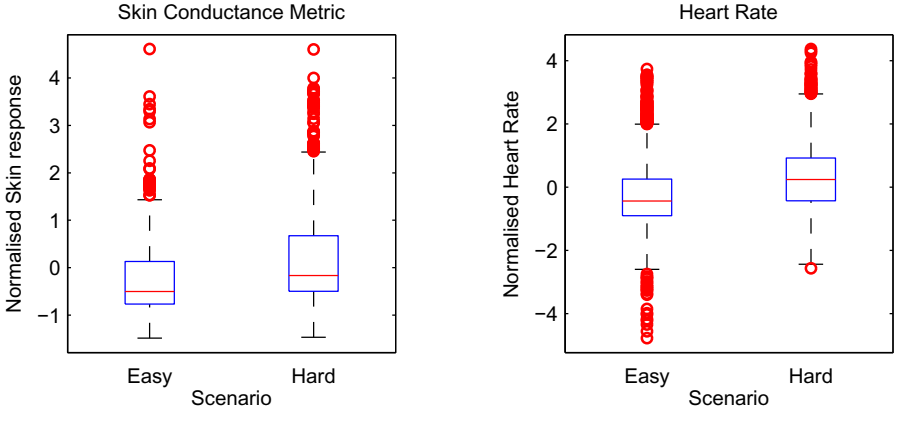
In order to test our hypothesis that participants were not sufficiently invested in their tasks to show signs of stress, we designed a second experiment. We retained the hardware setup of the driving simulator, but this time adapted it for remote control of a real vehicle: a Parrot AR.Drone quadricoptor [1]. This allowed us to maintain complete control over the experimental environment in a way not possible for on-road studies, but also gave our participants a real-world task that was not simulated.



(a) Best results from a single subject.



(b) Worst results from a single subject.



(c) Aggregate data from all subjects.

Fig. 2. Sample results from the simulated driving experiment. Although a small number of subjects showed good results, the aggregate data shows no effect.

3.1 Experimental Setup

The driving simulator described above was repurposed to allow remote control of the quad-rotor drone. This involved the addition of a joystick which allows more realistic control of a flying vehicle than the steering wheel and pedals used previously. The driving simulation software was removed and a live video feed from the forward-facing camera on the drone was projected onto the screen. Figure 3 shows the experimental setup. The drone itself was modified to operate in infrastructure wireless mode, allowing remote control throughout the building using our laboratory wireless network.

3.2 Subjects

Five subjects were recruited locally from within the university. All were 20 – 30 years of age. Some participants had previous experience flying the drone, while others were flying for the first time.

3.3 Procedure

On arrival, participants were instructed in the operation of the drone and given as long as they wanted to practise and get comfortable with the controls. This typically took around 10 minutes. Subjects wore sensors for measurement of skin conductance and blood volume pulse as well as a Dikablis Eye Tracker for measurement of absolute pupil position (which is not possible with EOG).

This experiment consisted of just two conditions, each interleaved and repeated several times. In the first ("easy") condition, participants were asked to fly the drone slowly from one end of a corridor to the other and back several times. In the second ("hard") condition, participants were asked to fly the drone down a corridor that included a 90° bend and then return to their starting point as fast as possible. In this condition they were also required to respond to an n-back secondary task. Single-digit numbers were read to them through a speech synthesiser and every time they heard a number they had to repeat the digit they heard two numbers previously. This type of secondary task has been widely used in previous studies designed to evoke cognitive overload and stress [7]. Whenever the drone crashed, it was reset in the corridor and the experiment continued. We considered the embarrassment of crashing the drone to be sufficient incentive for the subjects to avoid collisions. The experiment lasted approximately 40 minutes for each participant.

Fig. 3. The driving simulator, adapted for remote control of the Parrot AR.Drone

3.4 Data Analysis

The skin conductance data was bandpassed (0.001 Hz – 0.3 Hz) to remove high-frequency noise and long-term trends. Several features were calculated and then combined into a single "arousal" metric. Peaks were identified along with their preceding trough and then rise-rate and total height calculated. The product of these two features provides a metric which responds strongly to very sharp, tall peaks and largely ignores small, gentle slopes. These features are similar to those used by Healey and Picard in their previous study [5].

The blood volume pulse signal was bandpassed (0.5 Hz – 8 Hz) to remove noise and drift and scaled to mitigate the effects of changing circulation and sensor movement. These processing steps also removed variation between participants. A threshold was then applied to identify peaks corresponding to heart beats. The distribution of peak-heights in the BVP signal was plotted and a threshold chosen which would separate the tall peaks corresponding to heart beats from the others. Continuous beats-per-minute (BPM) were then calculated from the individual beat intervals and smoothed through a 15-beat moving average. The eye-movement data was recorded for future processing, but was not analysed in this experiment.

Both the skin conductance metric and the BPM were normalised for each participant by subtracting the mean and dividing by the standard deviation over the whole experiment. This allows us to compare participants with different physiological baselines on the same scale.

3.5 Discussion

Our intention was that the two scenarios would provide significantly different levels of cognitive load. The participants who could already fly the drone commented that the first condition was quite easy and that the second condition was extremely hard, validating our choice of scenarios for that group. The participants with less experience of the system found both scenarios rather difficult and some were unable to complete the harder of the two. As such, we only consider the results from subjects who completed both tasks. Figure 4 shows the results of the experiment. Readers will note the strong correlation between the heart rate, skin conductance and task difficulty in all individual subjects, and also in the aggregated data.

4 Recommendations

In this section we provide a set of recommendations for the design of experiments intended to measure the effects of varying cognitive load.

4.1 Location

Many previous studies have shown that varying cognitive load can cause measureable effects in drivers [5,10], pilots [12] and others. Most have performed data

collection in real-world scenarios where control of the experimental conditions is extremely coarse, such as driving at rush-hour versus driving in light traffic. These studies were essential for stimulating further work on this topic, but make quantification of the effects very difficult. Therefore, it is suggested that further, more controlled experiments are conducted in a laboratory setting. Although commercial moving-base flight simulators give an extremely realistic experience, they are too expensive for most research purposes, requiring an capital investment of around $10m and a recurrent cost of about $200 per hour.

4.2 Scenario

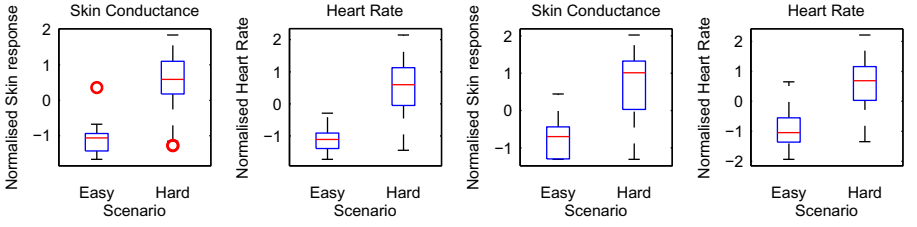
In order to elicit meaningful responses from subjects in a laboratory setting, it is essential that they are emotionally invested in the task undertaken. Entirely simulated scenarios may be suitable for studies of certain behaviours such as visual search [9,10], but if we expect to see genuine indicators of cognitive overload and stress, our subjects must really care about their performance. The approach described in our second experiment overcomes this by using a remote-control scenario where there is a real incentive for the participants not to crash the drone (i.e. it is expensive and could cause damage), but the environment is completely controlled.

4.3 Secondary Task

An important distinction that is rarely considered in the literature is between experiments where there are multiple tasks that are "easy" or "hard" (such as driving at rush-hour versus driving in light traffic) and experiments where a single primary task is performed, accompanied by a secondary task of varying difficulty. We have no reason to believe that the effects of a single task causing a high cognitive load will be equivalent to multiple simpler tasks. Consider the case of a driver in heavy traffic compared with the case of a driver in light traffic trying to operate a satellite navigation system. Both cases are important, but their effects may be very different. Our hybrid experimental environment will allow us to investigate both these types of cognitive load in detail.

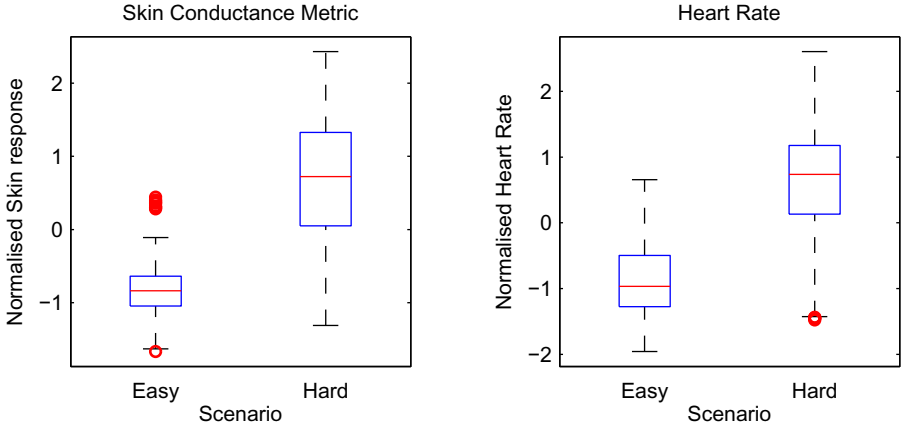
4.4 Data Collection and Analysis

When considering metrics for analysis, it is important to bear in mind potential applications. In particular, it is likely that most applications in control systems such as those described will require real-time data about the operator. Therefore, we should choose metrics that are calculable with minimal computation and as short a window of data as possible. All the metrics considered in this paper could be calculated continuously with a lag equal to the window size - generally no more than 10 seconds.



(a) Best results from a single subject.

(b) Worst results from a single subject.



(c) Aggregate data from all subjects.

Fig. 4. Results from expert pilots in the remote-controlled flying experiment. Compare to Figure 2, where the distinction between the scenarios in the aggregate data was much less clear.

5 Conclusion

Understanding the effects of cognitive load is an increasingly important area of research, and previous studies have shown that physiological effects are measurable in both laboratory-based and real-world scenarios. However, we argue that laboratory experiments with artificial tasks lack the external validity necessary for generalisation and data collection in real-world situations lacks the internal validity necessary for accurate quantification of the effects.

We have shown through the simulated driving experiment described in Section 2 that simulation of real-world scenarios is not sufficient to evoke consistent responses to cognitive load. Our remotely-controlled flying experiment (Section 3) confirms the hypothesis that this is due to a lack of emotional investment rather than the length of the scenarios or the effect of social interactions. We designed an experiment that provides subjects with a representative real-world task in an experimental environment that is completely controlled. Comparing

Figure 2 (c) with Figure 4 (c) we can see that the results of this experiment are significantly better.

Finally, we presented a set of experimental design recommendations for naturalistic data collection of the effects of operator cognitive load. We argue that, if these recommendations are followed, naturalistic physiological data collection will be possible in controlled experimental environments.

Acknowledgements. This work is supported by the EPSRC and Thales Research and Technology UK Ltd.

References

1. Parrot AR.Drone, `http://ardrone.parrot.com/`
2. Aasman, J., Mulder, G., Mulder, L.J.M.: Operator effort and the measurement of heart-rate variability. Human Factors 29(2), 161–170 (1987)
3. Backs, R.W., Seljos, K.A.: Metabolic and cardiorespiratory measures of mental effort: the effects of level of difficulty in a working memory task. International Journal of Psychophysiology 16(1), 57–68 (1994)
4. Hart, S.O., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research (1988)
5. Healey, J.A., Picard, R.W.: Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on Intelligent Transportation Systems 6(2), 156–166 (2005)
6. Healey, J.A., Seger, J., Picard, R.W.: Quantifying driver stress: Developing a system for collecting and processing bio-metric signals in natural situations. In: Proceedings of the Rocky Mountain Bio-Engineering Symposium (1999)
7. Jansma, J.M., Ramsey, N.F., Coppola, R., Kahn, R.S.: Specific versus nonspecific brain activity in a parametric N-back task. Neuroimage 12(6), 688–697 (2000)
8. Lisetti, C., Nasoz, F.: Affective intelligent car interfaces with emotion recognition. In: Proceedings of 11th International Conference on Human Computer Interaction (2005)
9. Recarte, M.A., Nunes, L.M.: Mental workload while driving: Effects on visual search, discrimination, and decision making. Journal of Experimental Psychology: Applied 9(2), 119–133 (2003)
10. Reimer, B., Mehler, B., Wang, Y., Coughlin, J.F.: The impact of systematic variation of cognitive demand on drivers visual attention across multiple age groups. In: Human Factors and Ergonomics Society Annual Meeting Proceedings, pp. 2052–2056 (2010)
11. Robinson, P., el Kaliouby, R.: Computation of emotions in man and machines. Philosophical Transactions of the Royal Society B: Biological Sciences 364(1535), 3441 (2009)
12. Veltman, J.A., Gaillard, A.W.K.: Physiological indices of workload in a simulated flight task. Biological Psychology 42(3), 323–342 (1996)