# Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks

**Conference Paper** · November 2012

DOI: 10.1145/2414536.2414602



**4 authors**, including:

Fang Chen
Jinan University (Guangzhou, China)
**275** PUBLICATIONS **4,087** CITATIONS
SEE PROFILE

Rafael A Calvo
Imperial College London
**384** PUBLICATIONS **11,022** CITATIONS
SEE PROFILE

# Using Galvanic Skin Response for Cognitive Load Measurement in Arithmetic and Reading Tasks

**Nargess Nourbakhsh**[1,2] **Yang Wang** [1] **Fang Chen** [1] **Rafael A. Calvo** [2]

[1] National ICT Australia (NICTA)
NSW 2015, Australia
{first.last}@nicta.com.au

[2] The University of Sydney
NSW 2006, Australia
{first.last}@sydney.edu.au

**ABSTRACT**
Galvanic Skin Response (GSR) has recently attracted researchers' attention as a prospective physiological indicator of cognitive load and emotions. However, it has commonly been investigated through single or few measures and in one experimental scenario. In this research, aiming to perform a comprehensive study, we have assessed GSR data captured from two different experiments, one including text reading tasks and the other using arithmetic tasks, each imposing multiple cognitive load levels. We have examined temporal and spectral features of GSR against different task difficulty levels. ANOVA test was applied for the statistical evaluation. Obtained results show the strong significance of the explored features, especially the spectral ones, in cognitive workload measurement in the two studied experiments.

**Author Keywords**
cognitive load, physiological signals, galvanic skin response, temporal analysis, spectral analysis

**ACM Classification Keywords**
H.5.2 Information interfaces and presentation: User interfaces - *Evaluation/methodology*.

**INTRODUCTION**
The term cognitive load is used to refer to the load that performing a particular task imposes on the person's cognitive system (Paas et al., 1994). It has extremely profound impacts on many aspects of human life including, but not limited to, learning (Sweller, 1994), safety in driving (Engstrom et al., 2005), aviation (Huttunen et al., 2011; Wilson, 2009), and user interface design (Saadé et al, 2007). Cognitive overload often leads to performance reduction and errors that in some cases such as air traffic control can have serious consequences. An adaptable system would be able to continuously monitor its users' experienced cognitive load and change its interactions with them when necessary to avoid cognitive overload. Such an adaptation obviously needs accurate and real-time mental load measurement.

Different methods have been applied for quantifying cognitive workload. Subjective method (Paas, 1992) or self-reporting is probably the commonest and, regarding implementation, the most convenient way. Nevertheless, it interrupts the natural routine of performing the principal tasks and also is not in real-time. Performance-based measurement techniques (Chandler et al., 1996) have been widely used as well but they still have the latter problem. Conversely, human behaviours - such as speech (Huttunen et al., 2011) - and physiological responses can reflect fluctuations of cognitive states over performing the intended task and without interfering with it.


*OZCHI'12*, November 26–30, 2012, Melbourne, Victoria, Australia.


Various physiological signals can be used for cognitive load measurement. So far, signals from heart, eye, brain, muscles and skin have been investigated in relation with cognitive load (Engstrom et al., 2005; Xu et al., 2011; Berka et al., 2007; Leyman et al., 2004; Wilson, 2009). Among different physiological signals, galvanic skin response (GSR), also referred to as electrodermal activity (EDA), is a low-cost, easily-captured, robust one. In this method the electrical conductance of the skin is measured through one or two sensor(s) usually attached to some part of hand or foot. Skin conductivity varies with changes in skin moisture level (sweating) and can reveal changes in sympathetic nervous system.

GSR has recently been investigated concerning mental status and emotions. (Nakasone et al., 2005) have successfully used skin conductance and muscle activity for emotion detection. In another study, skin conductance was measured to differentiate between a stress condition and a cognitive load condition, seeking the ability of detecting stress states (Setz et al., 2010). (Shi et al., 2007) also assessed GSR in stress and cognitive load situations and found correlations between readings of this signal and cognitive load. (Engstrom et al., 2005) found a weak effect of cognitive load on physiological signals including mean skin conductance. (Ikehara et al., 2005) evaluated GSR in relation with two levels of cognitive load. In contrast with other studies, they found skin conductance to decrease as task difficulty increases and explained it as a result of the easy task being tedious and too easy. (Wilson, 2009) analysed several physiological measures during different steps of flights and found out an increase in EDA response during take-off and landing which were expected to place the most cognitive demands on pilots. (Haapalainen et al., 2010) assessed mean, variance and median of GSR and other physiological signals against two cognitive load levels. They did not obtain any satisfactory results for GSR and explained that it might be related to the tasks type or their GSR sensors might not have been sensitive enough.

The abovementioned literature shows the usability of skin conductance in detecting emotions, but not many of the previous studies have found strong relations between

GSR features and mental workload. Moreover, to the best of our knowledge, none investigated spectral features and usually no more than one temporal feature has been observed in each study. Therefore, we aimed to perform a comprehensive study of GSR regarding cognitive load. In this paper, we have assessed GSR data captured from two experiments of different types each consisting of multiple cognitive load levels. We expected GSR amplitude and energy to have correlations with the amount of cognitive load. Hence time- and frequency-domain features have been examined against different task difficulty levels.

## FIRST EXPERIMENT: ARITHMETIC TASKS

### Data Collection

#### *Task Description*

This experiment included 8 arithmetic tasks with 4 difficulty levels. Each subject performed two trials of each task level and the whole eight trials were performed in a randomised order. First to fourth difficulty levels respectively included binary numbers (0 and 1), one-, two- and three-digit numbers. In each task four numbers were shown one by one, each for three seconds. Subjects were asked to add these four numbers and select the correct answer from three numbers which were next presented. Trials with a same difficulty level included different numbers. Before appearing the first number of each task, one to three 'x' symbols (according to the number of digits in the task) were presented for three seconds. There was no time limit for answer input.

#### *Apparatus*

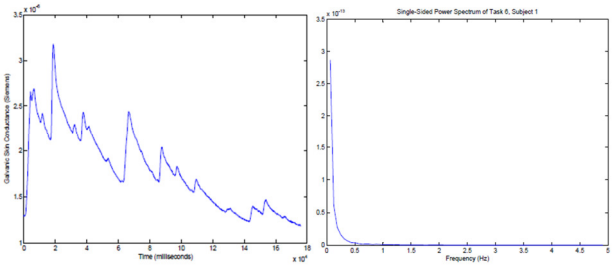
To collect galvanic skin response, the GSR device from ProComp Infiniti of Thought Technology Ltd was used and the sensors were attached to the subjects' left hand finger. The sampling frequency was 10Hz. A 21" LCD monitor and a usual computer mouse were peripherals for interaction between participants and a PC running the tasks. Another PC collected the signals through GSR sensors and was synchronised with the first one.

#### *Participants*

Twelve 24 to 35-year-old male volunteers participated in this experiment. They signed a consent form before the experiment and were awarded with a $10 movie voucher for their participation. The experiment was approved by Human Research Ethics Committee of the University.

### Methods

In both experiments, we observed that the GSR values are highly subjective, that is they differ from person to person. When the recorded values are analysed without any pre-processing, no significant results can be obtained. We hypothesised that some sort of normalisation which omits this dependency on subjects might result in useful and informative findings. As will be discussed in the following sections, we scaled the data of all subjects in a same range by using a normalisation and this hypothesis was confirmed by our results. We assessed time and frequency domain features, and averaged each feature between tasks with same difficulty levels for each subject. One-way ANOVA test was applied to statistically evaluate each temporal/spectral feature. Figure 1 shows the GSR signal of one of the subjects and the spectral curve of one task of that subject.




**(a)** **(b)**

**Figure 1. (a) Galvanic Skin Response of Subject 1. (b) Single-Sided Spectrum of Task 6, Subject 1.**

#### *Time Domain Features*

We calculated the summation of GSR values (accumulative GSR) from *appearing the first number until inputting the answer*. We refer to this period as *task time*. In addition, we averaged the GSR value of each task over task time and assessed this feature as well. In order to omit the subjective differences, we normalised each participant's data by dividing task signal by the mean value of all tasks of the subject:

$$Normalised\_GSR(i,k,t) = \frac{GSR(i,k,t)}{\frac{1}{m}\sum_{j=1}^{m}\sum_{t=1}^{Tij} GSR(i,j,t)} \quad (1)$$

$GSR(i,k,t)$ is value of each data-point at time $t$ of task $k$ of subject $i$ and $m$ is the number of tasks (in this experiment $m$=8). We calculated accumulative GSR and average GSR for task $k$ of subject $i$ as below:

$$Acc\_GSR(i,k) = \sum_t \quad Normalised\_GSR(i,k,t) \quad (2)$$

$$Avg\_GSR(i,k) = \frac{\sum_t \quad Normalised\_GSR(i,k,t)}{T} \quad (3)$$

where $T$ is a whole task time and $t$ is a sampling time.

#### *Frequency Domain Features*

In frequency domain we examined different variations of *power spectrum*:

$$P(\omega) = \frac{1}{N} Y(\omega) Y^*(\omega) \quad (4)$$

where $P$ is power spectrum, $\omega$ is angular frequency, $N$ is the length of signal, $Y$ is frequency spectrum and $Y^*$ is the complex conjugate of $Y$. Power feature was observed to have non-zero values in frequencies less than 1Hz, mostly less than 0.5Hz (Figure 1(b)).

For each subject, power spectrum of each task (both as a whole and cut into frames) were calculated. For the whole tasks, we have normalised each task by mean of all tasks of the subject. In segmentation, we have divided each task into frames of 16, 32 and 64 data-points length, calculated the power spectrum over each frame and normalised each frame $k$ of any task $j$ of each subject $i$ by dividing the average power value of the frame ($power\_frame(i,j,k)$) by average power of all frames of the particular subject:

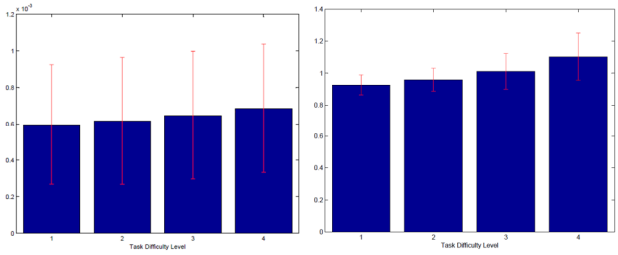
$$N\_power\_frame(i,j,k) = \frac{power\_frame(i,j,k)}{\frac{1}{m}\sum_{l=1}^{m} power\_frame(i,j,l)} \quad (5)$$

where $m$ is the total number of frames of subject $i$. The whole procedure was performed four times for each task of each subject, once on the whole task's data and once for every segmentation. In each run, power features of direct current (DC, frequency=0), the whole frequency range (excluding DC part), and below 1 Hz (as this range had most non-zero power values) were calculated.

## Results

### *Time Domain*

As mentioned before, GSR values are subject-depended, i.e. different people have different GSR value levels. This can be observed from Figure 2(a). Although there is an increasing trend between task difficulty and average sum feature, the huge standard deviations reveal how large the between-subjects difference for each task is. The results of statistical analysis (ANOVA test) of the accumulative GSR without any normalisations were insignificant (p=0.9162). We scaled data of all subjects in a similar range and analysed the normalised data. The effectiveness of this normalisation can be observed from Figures 2(b): the increasing trend of the feature versus task difficulty levels exists and the value for each difficulty level has a much smaller standard deviation.




**(a)** **(b)**

**Figure 2. Mean and Standard Deviation of Accumulated GSR Values for All Subjects. (a)Without Normalisation. (b)With Normalisation.**

This normalisation leads to highly significant results in respect to differentiating between the four task difficulty levels (F=7.21, p=4.86x10$^{-4}$). It can be observed that accumulative GSR produces results which are a lot more significant than those produced by average GSR (with normalisation: F=2.43, p=0.0790; without normalisation: F=0.02, p=0.9973).

### *Frequency Domain*

In frequency domain, same as time domain, without normalisation no significant results could be obtained. Tables 1 and 2 show the results of statistical analysis of subjective-normalised frequency domain features (NDP: number of data-points in each segment).

| Whole Task | NDP=16 | NDP=32 | NDP=64 |
|---|---|---|---|
| F=4.96<br>p=0.0047 | F=3.22<br>p=0.0316 | F=3.14<br>p=0.0346 | F=3.08<br>p=0.0371 |

**Table 1. ANOVA Test Results of Average GSR Frequency Power of DC Part in First Experiment**

It can be seen from Table 1 that the whole task and all the segmentations produce significant results; however the best is obtained from using the whole task. On the other hand, a slightly different pattern is observed regarding segmentations. Although these results are very close, smaller frames result in a little more significance.

| Whole Task | NDP=16 | NDP=32 | NDP=64 |
|---|---|---|---|
| F=3.11<br>p=0.0358 | F=5.36<br>p=0.0031 | F=5.14<br>p=0.0039 | F=4.65<br>p=0.0066 |

**Table 2. ANOVA Test Results of Average GSR Frequency Power (Excluding DC) in First Experiment**

Table 2 reveals that the smaller the frame length is, the more significant the results are. The best results, which are highly significant, belong to 16-datapoint-length frames. The results shown in table 2 were almost repeated when analysing the frequencies below 1 Hz. This can be explained with the observation mentioned before: the most non-zero values of the signals relate to frequencies below 1 Hz.

## SECOND EXPERIMENT: READING TASKS

## Data Collection

### *Task Description*

Three silent reading tasks were performed. Each task consisted of four slides of text and each slide was presented for 30 seconds. The participants were supposed to find words of certain lengths in each slide. There were three task difficulty levels: the easiest task required finding three-letter words; in the medium task subjects should find three- and four-letter words; in the hardest task they were supposed to find three-, four- and five-letter words. They were asked to click on the left, middle or right mouse button when finding a three-, four- or five-letter word respectively. Task difficulty orders were randomly chosen.

### *Apparatus*

The galvanic skin response was recorded using the GSR device from ProComp Infiniti of Thought Technology Ltd and the sensors were attached to the subjects' left hand finger. The sampling frequency was 10Hz. A 19" LCD monitor and a usual computer mouse were peripherals for interaction between participants and a PC running the tasks. Another PC collected the signals through GSR sensors and was synchronised with the first one.

### *Participants*

Twelve 16 to 46-year-old male and female volunteers participated in this experiment. Each participant was awarded with a chocolate bar as an added incentive to their participation. Before performing the experiment, each subject signed a form allowing the experimenters to collect and use the data for research. The experiment was approved by Human Research Ethics Committee of the University.

## Methods

In this experiment we almost repeated the analysis performed on the first experiment data and calculated the features using the same formulas. However, there were some differences. As mentioned before, our reading experiment consisted of three tasks of different difficulty levels for each participant, every task level was performed once by each subject, and the duration of the tasks were longer than the arithmetic tasks. We assessed time and frequency domain features and applied one-way ANOVA test to statistically evaluate each time/frequency domain feature.

### *Time Domain Features*

We calculated the *accumulative GSR* during task time and also *averaged the GSR* value of each task over task time. A similar normalisation was performed: dividing data of each task of each subject by mean of all tasks of that subject.

*Frequency Domain Features*

For each subject, spectral features of each task were calculated. The analysis was performed for the whole tasks and different segmentations. For the whole tasks, we have normalised each power spectrum by average frequency power of all tasks of the subject. In segmentation, we have divided each task into frames of 16, 32, 64 and 128 data-points length, calculated the power spectrum over each frame and normalised each frame of any task by dividing the power value of that frame by average of all frames of all tasks of that particular subject. The average frequency power was examined for DC part and the whole frequency range excluding DC.

## Results

*Time Domain*

Same as the first experiment and due to the subjective nature of GSR values, temporal features of this experiment could not make significant results before being normalised (F=0.24, p=0.7868 for accumulative GSR; F=0.09, p=0.9676 for average GSR). The normalisation produced significant results for average GSR (F=2.98, p=0.0444) and improved accumulative GSR results (F=1.66, p=0.1925).

*Frequency Domain*

Due to the subjective nature of GSR values, no significant results were obtained before normalisations were applied. Tables 3 and 4 show the results of statistical analysis of normalised spectral features. The whole tasks lead to the most significant results when assessing the frequency spectrums in DC part of the signal, and the segmented signals produce better results when the DC part is excluded. This is consistently with the findings of the first experiment.

| Whole Task | NDP=128 | NDP=64 | NDP=32 | NDP=16 |
|---|---|---|---|---|
| F=8.43 p=0.0002 | F= 3.7 p=0.0203 | F=3.51 p=0.0249 | F=3.52 p=0.0246 | F=3.43 p=0.0269 |

**Table 3. ANOVA Test Results for Power Feature of DC Parts in Second Experiment**

| Whole Task | NDP=128 | NDP=64 | NDP=32 | NDP=16 |
|---|---|---|---|---|
| F=0.15 p=0.8652 | F=4.76 p=0.0067 | F=3.98 p=0.0152 | F=2.19 p=0.1059 | F= 1.54 p=0.2200 |

**Table 4. ANOVA Test Results for Power Feature (Excluding DC) in Second Experiment**

## CONCLUSION

GSR is a nonintrusive easily-captured physiological signal which is being explored as one of cognitive load measures. In this study, we investigated different time- and frequency-domain features of GSR in multiple difficulty levels of arithmetic and reading experiments. A normalisation was applied to omit the subject-dependency of GSR data. The results show that normalisation effectively improves the significance of distinction between the cognitive load levels for mean and accumulative GSR and the spectral features. In frequency domain, we investigated the behaviour of power spectrum of the whole tasks as well as smaller segments within each task, in DC part and the rest of the frequency range. Regarding to DC part, the best results belonged to the whole tasks; however, when analysing the frequencies below 5Hz or below 1Hz, segmented signals produce more significant results in both experiments. Findings of this research suggest the usability of the explored features in future systems where measuring cognitive load would result in smart and adaptable interactions with humans. Our future work will include applying machine learning techniques and assessing the performance of other physiological features in cognitive load detection.

## ACKNOWLEDGMENTS

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

## REFERENCES

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., Olmstead, R. E., Tremoulet, P. D., Craven, P. L. EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks. Aviation, Space, and Environmental Medicine, 78, 5 (2007), B231-44.

Chandler, P., Sweller, J. Cognitive Load While Learning to Use a Computer Program. Applied Cognitive Psychology, 10, 2 (1996), 151–170.

Engstrom, J., Johansson, E., Ostlund, J. Effects of Visual and Cognitive Load in Real and Simulated Motorway Driving. Transportation Research Part F, 8, 2 (2005), 97-120.

Haapalainen, E., Seungjun, K., Forlizzi, J. F., Dey, A. K. Psycho-Physiological Measures for Assessing Cognitive Load, *In Proc. Ubicomp* 2010, ACM Press (2010).

Huttunen, K., Keränen, H., Väyrynen, E., Pääkkönen, R., Leino, T. Effect of Cognitive Load on Speech Prosody in Aviation: Evidence from Military Simulator Flights. Applied Ergonomics, 42 (2011), 348-357.

Ikehara, C. S., Crosby, M. E. Assessing Cognitive Load with Physiological Sensors. In Proc. HICSS 38 (2005), 295a - 295a.

Leyman, E., Mirka, G., Kaber, D., Sommerich, C., (2004). Cervicobrachial Muscle Response to Cognitive Load in a Dual-Task Scenario. Ergonomics, 47(6), 625-45.

Nakasone, A.; Prendinger, H. & Ishizuka, M. Emotion Recognition from Electromyography and Skin Conductance, In Proc. BSI 2005 (2005), 219-222.

Paas, F. G. W. C. Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics: A Cognitive-Load Approach. Journal of Educational Psychology, 84, 4 (1992), 429-434.

Paas, F. G. W. C., Van Merriënboer, J. G. Instructional Control of Cognitive Load in the Training of Complex Cognitive Tasks. Educational Psychology Review, 6, 4 (1994), 351-371.

Saadé, R. G. & Otrakji, C. A. First impressions last a lifetime: effect of interface type on disorientation and cognitive load. Computers in Human Behaviour, 23, 1 (2007), 525-535.

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tr¨oster, G. Discriminating Stress from Cognitive Load Using a Wearable EDA Device. Technology, 14, 2 (2010), 410-417.

Shi, Y., Ruiz, N., Taib, R., Choi, E. H., Chen, F. Galvanic Skin Response (GSR) as an Index of Cognitive Load. In Proc. CHI 2007 Work-in-Progress, (2007), 2651-2656.

Sweller, J., Cognitive Load Theory, Learning Difficulty, and Instructional Design. Learning and Instruction, 4, 4 (1994), 295–312.

Wilson, G. F. An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures. The International Journal of Aviation 12, 1 (2009), 3-18.

Xu, J., Wang, Y., Chen, F., Choi, H., Li, G., Chen, S., Hussain, S. Pupillary Response Based Cognitive Workload Index Under Luminance and Emotional Changes. In Proc. CHI 2011, ACM Press (2011), 1627-1632.