Preliminary version of: Khan, I. A., Brinkman, W. P., & Hierons, R. (2013). Towards estimating computer users' mood from interaction behaviour with keyboard and mouse. *Frontiers of Computer Science*, *7*(6), 943-954.

Towards Estimating Computer Users' Mood from Interaction Behaviour with Keyboard and Mouse

Abstract The purpose of this exploratory research was to study the relationship between the mood of computer users and their use of keyboard and mouse to examine the possibility of creating a generic or individualized mood measure. To examine this, a field study (n = 26) and a controlled study (n = 16) were conducted. In the field study, interaction data and self-reported mood measurements were collected during normal PC use over several days. In the controlled study, participants worked on a programming task while listening to high or low arousing background music. Besides subjective mood measurement, Galvanic Skin Response data was also collected. Results found no generic relationship between the interaction data and the mood data. However, the results of the studies found significant average correlations between mood measurement and personalized regression models based on keyboard and mouse interaction data. Together the results suggest that individualized mood prediction is possible from interaction behaviour with keyboard and mouse.

Keywords: keyboard, mouse, interaction, mood measure, computer users, programming

1 Introduction

With the aim of enhancing the interaction with computer systems by considering features such as recognition, interpretation and expression of moods and emotions [1, 2], affective computing has received a considerable amount of research attention in the last decades [3]. A computer that recognizes affects (moods/emotions) is proposed to be more effective and natural [4]. Such computers have the promise of making tasks easier to do, and providing effective support to reduce frustration levels [5]. Similarly such computers might provide encouragement and comfort [5]. This seems desirable as research shows that computer users do get frustrated and moody [6]. For example, Ross & Zhang [7] showed that programmers get frustrated and moody because of deadlines, pressure to produce error free programs and pressure to produce some useful output. Furthermore, their performance on tasks such as debugging has been shown to be affected by their mood [8]. To develop systems that respond appropriately to users' situation, it is therefore important that computers can estimate the users' mood.

The task of measuring the user's mood however is not a trivial one. Several methods have been suggested (see related work section). However, they all have the drawback of requiring some kind of additional, often expensive equipment, or obstruct or interrupt the primary task the user is engaged in [4]. This paper therefore argues for an alternative way of measuring users' mood, by analysis of their keyboard and mouse behaviour. The two empirical studies presented here show that individualized mood measure based on user interaction is possible, while no support was found for a generic measure. These new insights are the main scientific contributions of this paper. A mood measure such as suggested here could potential benefit for example mental health applications, or training applications. Before presenting these studies, the next section describes related work on measuring computer users' mood.

2 Related Work

The task of affect recognition is complex. Various affect recognition methods like self-reports [9], physiological methods [10] and automatic facial recognition [11] are in practice. One of these methods also includes affect recognition and prediction from the individual's behaviour. Affects are known to have an impact on behaviour. As moods last for a longer time, their influence on behaviour might be more prominent than emotions [4]. The impact of moods may either be on informational behaviour or directive behaviour [12]. The informational impact of moods affects judgements, which then will results in behavioural adjustments while a directive impact influences behavioural preferences to fulfil hedonic[1] motives [12]. Moods also affect behaviours like inter-group discrimination [13], helping [14], judgments and evaluative judgments [15], helping tasks that are incompatible with good moods [16], acceptance of certain level of risks [17], decision rules in risk situations [18] and decision making [19].

Although significant amounts of literature confirm the impact of moods on behaviour, there is limited literature on mood measuring techniques from behaviour. One of the exceptions is Johnson et al. [20] who acquired ordered sets of training data from human interactions. These interactions included shaking of hands and body movements that are also called spatial interactions. Various other techniques are in practice to measure moods like automatic facial recognition [21], human movement recognition [21], emotion recognition from speech [22] and action recognition [23]. A psychophysiological measure like electroencephalograms (EEG) or electromyograms (EMG) can also be used to study unconscious functions of human body e.g. affects [24].

There is much research which shows an increase or decrease of various psychophysiological measures like

[1] Living and behaving in ways that mean you get as much pleasure out of life as possible

heart rate, respiration rate, finger pulse volume amplitude, eye blink rate and finger temperature under conditions like strain [25]. There is also field, simulation and laboratory research available on the effects of arousal on psychophysiological measures. For example, Chanel et. al., [26] used EEG to measure valence and arousal in emotion recall conditions. Various researches like Boucsein & Thum [27] focused on bio-signals from skin to determine emotional aspects. They found an increase in NS.SCR (non-Specific Skin Conductance Response) frequency while examining emotional strain during prolonged interruption of patent examiners.

Electrodermal activity (EDA) is considered to be the most convenient measure taken from the skin and is measured as SCR or as SRR (Skin resistance response) [25]. Researchers found psychophysiology parameters like increase in heart rate (HR) and respiration rate related to human-computer interaction [28] and computer related data entry work [29]. However skin response is also used widely for studying emotional responses related to computers. For example Sakurazawa et. al. [24] utilized galvanic skin responses in computer gaming. Lissetti and Nasoz [30] proposed an interface to sense a user's emotional and affective states by using visuals (images, videos), Kinaesthetic (Automatic nervous system signals) and Auditory (speech).

All methods reviewed above need complex software and algorithms or dedicated hardware like video cameras, pressure sensitive devices and other physiological devices [4]. These methods can also be a source of disruptions in user attention. One of the alternatives could be to analyse keyboard and mouse use. This would reduce overheads as they are the most basic devices used with computers. Computer users are familiar with these devices and there is no risk of external device disruptions. These devices are cheap as they are available with almost every computer and implementation of the needed functionality is also not difficult [4].

Very limited literature focuses on the possibility of estimating moods from the users' interactions with keyboard and mouse. Some examples include Mahr et al. [31], who used mouse motions to detect emotions with some significant correlations with arousal. However, they found no correlation between mouse motions with valence. Another example is Zimmerman et al. [4] who used keyboard and mouse and stored keystrokes and mouse movements in a log file to find correlations of these events with affective states. Zimmerman et al. [4] proposed experiments on the possibility of mood estimation from keyboard and mouse usage, and in 2010, Khanna and Sasikumar [32] reported on such a kind of experiment. They found that typing speed decrease when their participants were in a negative emotional state, and increased when they were in a positive emotional state. Others like [33] have investigated the possibility of using keyboard usage to improve visual-facial emotion recognition.

The two studies presented in this paper are a continuation of this exploration into the relationship between mood and keyboard, mouse behaviour. An argument is made that although a generic mouse-keyboard based mood measure is probably difficult to establish, an individualized mood measure is possible.

Russell [34], Bradley & Lang [35], and Mehrabian [36] argue that affective stimuli complexity and meaning can be described by three basic emotional dimensions: valence, arousal, and dominance. Valence and arousal are the primary dimensions, as they account for most of the variance in affective reactions [35], and are therefore important components of any mood measurement including the one presented in this paper.

3 Study 1 – Field Study

3.1 Experimental Material

In the first study keyboard and mouse events were recorded in log files over several days. In addition, at fixed intervals self-reported moods of the participants were also obtained and recorded in log files with the help of custom built background application. This application also presented a mood rating dialogue box that appeared after every twenty minutes. The purpose of the mood rating dialogue box was to record participants self-reports of their moods on the valence and arousal scale called SAM (Self-Assessment Manikin) measurement devised by Lang [37]. SAM represents valence and arousal along a nine-point scale using graphical characters. Arousal scale ranges from manikin with excited open eyes (high arousal) to sleepy closed eyes (low arousal). Similarly, the valence scale ranges from smiling happy figure (high valence) to frowning unhappy figure (low valence).

In addition the application provided various other functionalities such as pause logging[2] for 5 and 10 minutes, pause logging for variable time in minutes, exit logging[3], and stopping the pop-up[4] dialogue box. Participants were also able to uninstall the application with all their data deleted, if they wanted to withdraw from the study.

Categories of the keystrokes recorded in the log files were: (1) capital alphabets recorded as 'capital alphabet', (2) lower-case alphabets recorded as 'short alphabet', (3) numbers recorded as 'numeric', and (4) some special character like /, @ etc. were recorded as 'special character'. Keystrokes were only recorded in categories to

[2] It was assured to participants that the application would not record any personal information. To further increase the trust, pause logging for specific time interval functionality was provided

[3] Participants were able to exit logging at any time. The application was able to restart itself at the next computer boot up; Participants were also able to restart application without rebooting the computer.

[4] The appearance of a pop-up dialogue box after a fixed interval could be annoying especially when doing very important and concentration demanding work. This functionality was provided to stop appearance of pop-up dialogues until the participant restart it again.

provide protection to participants' personal data and information.

Table 1: Data logged

Data type	Description
Window name	This data was used to identify the application that was in use at specific event time. However special care was taken not to identify the names of the documents that participants were working on. The application was developed to record only specific application names like Microsoft Word, Internet Explorer, Visual Studio etc.
Keyboard or mouse event	This data was used to identify a particular event. An event could either be a mouse clicks or a key press. Mouse click events were identified as "Mouse button". Key press events were further divided into two categories i.e. of Key Up and of Key Down.
Date and time of event	This was used to record date and time at which event occurred.
Category of event	This was used to store the category of the event that occurred. It stored "Left" or "Right" if a mouse clicked and stored "Short Alphabet", "Capital Alphabet", "Numeric key", "Special Character" etc. if the event was a key press event.

3.2 Participants

Participants were 26 frequent computer users. Their average age was 27 years (*SD* = 3). Their age range was 22 – 34 years. Fifty percent of the participants classified themselves as programmers, 42% as expert computer users, and 8% as medium computer users. Participants mean experiences with computer was 5 years (*SD* of 1.6) with a range of 2-9 years. There were only two female participants in this study.

3.3 The Experiment Setup

The department ethics committee approved this experiment. The application was installed on participants' computers as a background running process. Participants were asked to sign a consent form before they could use this application. Four different kinds of data were recorded in the log files as is shown in Table 1. As explained earlier, this application also recorded self-reported moods using the mood rating dialogue in the log files.

3.4 Results

3.4.1 Preparation of Data

Participants were instructed to answer at least 86 mood-rating dialogues to have 80% probability of finding at least medium size correlations [38]. The background application recorded events for (on average) 8 days. A single log file was created for each day, and these files were later merged into a single log file. Each participant log file contained an average of 0.1 million lines of event recordings. An application was developed to extract the required information from the log files. The application extracted self-reported arousal and valence values from these log files and keyboard/mouse behaviour within 10 minute windows around these mood ratings as shown in Fig 1.

The basic measures taken for each window were:

- Self-reported valence and arousal values that a participants provided
- Total number of events around a particular mood rating
- Average time between these events
- Total windows switched
- Standard deviation of the time between events
- Number of backspace and delete key events
- Number of alphabetical and numerical key events
- Number of mouse clicks
- Number of all other keys

These measures are termed as log variables in the rest of the text. Slots with 10 or fewer events were removed as interaction was considered too limited for analysis. Another filter was based on the key press and mouse click rates. All the key press and mouse click events with less than 50 milliseconds difference from the previous event were filtered. This value was chosen since, while a novice has a typing speed of 1000 milliseconds, a champion typist has an average speed of 60 milliseconds to type a character [39]. A difference less than or equal to 50 milliseconds might be an indication that a participant was stroking and clicking haphazardly. Similarly all the key press and mouse click events with more than 20,000 milliseconds (20 seconds) difference from the previous events were also filtered because people waiting more than 20 seconds to type a key press might be involved in some other activity such as reading a website or other documents. These filters, however, reduced the initial 86 samples per individuals to a range of 83 to 21 samples per individual.

3.4.2 Analysis

As data was collected from multiple individuals over multiple events, a meta-analysis approach [40] was taken. This meant that first Pearson correlations were calculated between interaction data in ten minutes window and the valance or arousal ratings of each individual. Next these correlations were combined into a weighted average correlation across the individuals. The weighting was based on the sample size of the individual correlations. As the individual correlations were obtained from a single individual, the sample size of the meta-analysis was set to 26, the number of participants. The setup of the experiment was similar for each participant; therefore a fixed-effect model was used to compute summary effect.

To examine the possibility of a generic interaction based mood measure, *weighted correlations* and a *Fisher's z score (i.e. effect size)* were computed from the calculated correlations of each participant's valence, arousal ratings and the interaction measures. Table 2 shows the respective weighted correlations, Fisher's *z* transformations and the associate *p* values. No weighted correlations between the interaction data and mood rating reached a significant level (*p*. > 0.05). These results therefore provide no support for the idea of a generic mood measure based on interaction data.

The next step of the analysis was to examine whether an individualized interaction based mood measure is possible. For each participant a multiple regression analysis was conducted with either the valence or the arousal rating as the dependent variable and the eight interaction measures as independent variables by using the Enter method. The *R*-values of these individual analyses were again combined by calculating a weighted average correlation across the participants.

.

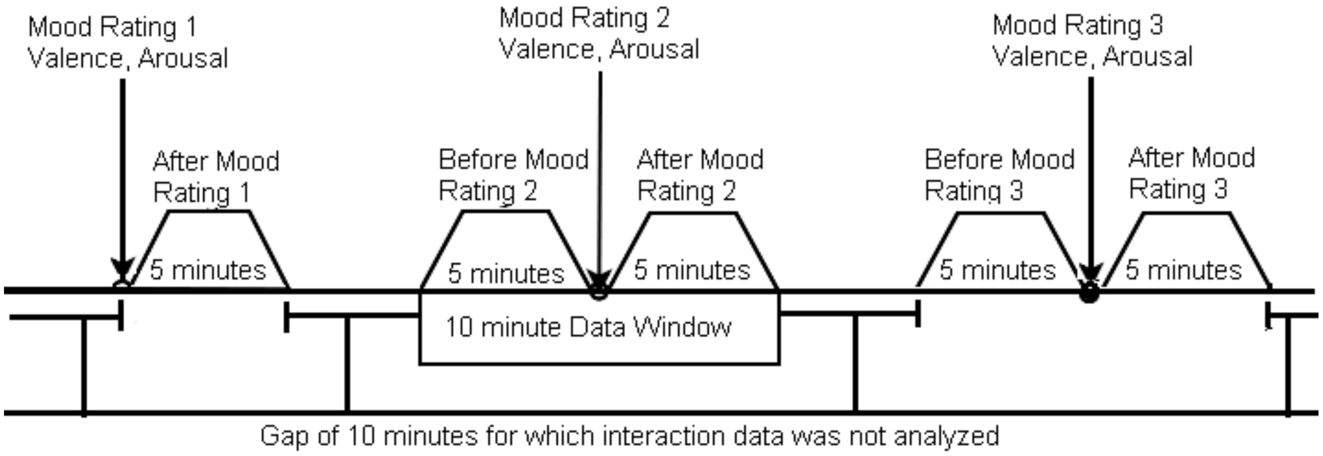

Fig 1: Time windows used to take events around self-reported valence/arousal.

Table 2: Weighted correlations, Fisher's *z* scores and *p* values of valence arousal ratings with behavioural variables

	Valence							
Measures	Events	Average Time Between Events	Windows Switched	SD Time b/w Events	Back space, delete Keys	Alphabetic numeric keys	Mouse Clicks	Others keys
Weighted r	-0.014	0.0044	0.030	-0.009	-0.01	0.025	-0.065	-0.002
Fisher's *z*	-0.013	0.0041	0.030	-0.01	-0.01	0.025	-0.066	-0.002
p value	0.52	0.49	0.44	0.51	0.52	0.44	0.63	0.50
	Arousal							
Weighted r	-0.06	0.034	-0.021	0.015	-0.059	0.006	-0.068	-0.046
Fisher's *z*	-0.06	0.035	-0.022	0.015	-0.059	0.006	-0.070	-0.047
p value	0.62	0.43	0.54	0.46	0.62	0.48	0.63	0.59

Table 3 shows results for the valance prediction with a weighted correlation of 0.36 and significant (Z = 1.97, one-tailed[5] *p*. = 0.025) Fisher's *z* score of 0.39. This suggests that on an individual level there was a significant medium size [38] relation between people's interaction behaviour and their valence rating. A similar observation could be made for arousal rating (Table 4). The weighted average correlation was 0.35, i.e. a medium size effect [38], with a significant (Z = 1.91, *p*. = 0.028) Fisher's *z* conversion of 0.37. This suggests that arousal of an individual might also be predictable from its interaction behaviour with keyboard and mouse.

3.4.3 Discussion

The results of study 1 provide clear support for an individualized mood measure based on user interaction behaviour. Interesting in this study is also what was not found. No support was found for a generic measure based on user interaction behaviour. The ecological validity, i.e. the study closely approximates real-world situations, suggests that these findings are based on natural mood variations and therefore applicable to everyday working situations. To strengthen these conclusions even further, the next step was therefore to reflect on how a second confirmation study could address potential counterarguments, i.e. potential limitations of study 1, as any empirical study has limitations. Two limitations could be identified as alternative explanations why no generic correlations were found. They were: (1) Participants' mood ratings had a small standard deviation for both valance (M = 4.80, SD = 0.94) and arousal (M = 5.19, SD = 1.15) with both means near to 5, the centre point of both ratings. This suggests that participants' moods

[5] The correlation results of regression analysis cannot be negative, a one-tailed test is therefore appropriate in this case.

variation was limited and might therefore provide only limited data to explore correlations. This might be a side effect of a field study, which has no experimentally controlled mood induction to elicit more extreme mood variance. (2) Also participants cancelled 27% of their mood rating dialogues on average. Two participants reported them as a distraction. Participants might therefore especially have cancelled them when in either high arousal or low arousal situations.

Given these two limitations, the second study was therefore setup to have more mood variance and no potentially distracting mood rating dialogues.

4 Study 2 – Controlled Study

4.1 Experimental Material and Design

The basic design of the second study was similar to the first study; however this time the variation in mood was systematically manipulated by using music. The keyboard and mouse events were again recorded in a log file. The self-reported moods measurement, with its problems concerning distraction or missing reports, was replaced by galvanic skin response (GSR) measurement. As mentioned before, mood inducing music was used in the background to induce moods. The aim of the study was to find correlations between keyboard and mouse events and GSR measurements. Measuring valence and arousal without human intervention was the basic aim however previous research [41] indicated that it is difficult to accurately measure valence information from physiological sensors. Therefore this experiment only targeted arousal for objective measurements. To confirm the validity of mood inducing music, first the impact of mood inducing music was analysed based on GSR ratings. To handle various external mood-affecting factors, like the nature of the task, programming skills, etc., only mood induction via music was systematically changed between conditions. The other external factors could therefore be assumed to have affected the experimental conditions equally, or more precisely, not systematically biased the results in one condition. Any difference observed in the physiological data above the noise level, caused by individual difference of the participants (e.g. programming skill, preference to listen to music etc.), could therefore be attributed to the only systematically varied environmental factor, i.e. the music. For this a 2x2 factorial design was applied with arousal (High and Low) and music type (Based on the literature vs. Brought by participants) as independent variables. GSR measurements were used as dependent variable. GSR meter is a psychometric research biofeedback monitor designed to measure the electrical conductivity of the skin. The changes are a response to certain emotional reactions. Increase in the activity (high arousal, high valence) results in the increase of electrical conductivity whereas relaxations (low arousal, low valence) results in the decrease of electrical conductivity [43]

A GSR meter was used to measure GSR by attaching electrodes to the base of the left-hand middle finger and ring finger. In addition keyboard key press and mouse clicks were also recorded like study 1, except that there was no self-reporting of mood.

Table 3: Correlations and effect size of regression analyses on individual sample set to predict Valence rating

Participant	Individual sample size	*r*	Fisher's *z*
1	73	0.40	0.42
2	49	0.51	0.56
3	72	0.39	0.41
4	68	0.33	0.35
5	46	0.31	0.32
6	67	0.22	0.22
7	19	0.43	0.46
8	54	0.41	0.43
9	56	0.60	0.69
10	18	0.43	0.46
11	69	0.44	0.47
12	65	0.23	0.23
13	30	0.40	0.43
14	62	0.30	0.31
15	66	0.28	0.28
16	69	0.20	0.20
17	46	0.45	0.48
18	57	0.27	0.28
19	45	0.64	0.76
20	31	0.50	0.54
21	46	0.31	0.32
22	58	0.47	0.51
23	59	0.27	0.27
24	66	0.29	0.30
25	32	0.45	0.48
26	68	0.31	0.32
n meta-analysis		*weighted r*	*weighted Fisher's z*
26		0.36	0.39

Table 4: Correlations and effect size of regression analyses on individual sample set to predict Arousal rating.

Participant	Individual sample size	*r*	Fisher's *z*
1	73	0.40	0.42
2	49	0.51	0.57
3	72	0.28	0.29
4	68	0.26	0.27
5	46	0.40	0.42
6	67	0.24	0.25
7	19	0.59	0.68
8	54	0.24	0.25
9	56	0.35	0.36
10	18	0.66	0.78
11	69	0.50	0.55
12	65	0.35	0.36
13	30	0.52	0.58
14	62	0.35	0.36
15	66	0.23	0.23
16	69	0.28	0.28
17	46	0.34	0.36
18	57	0.31	0.32
19	45	0.39	0.42
20	31	0.59	0.67
21	46	0.33	0.35
22	58	0.37	0.39
23	59	0.30	0.31
24	66	0.47	0.51
25	32	0.13	0.13
26	68	0.33	0.35
n meta-analysis		*weighted r*	*weighted Fisher's z*
26		0.35	0.37

4.1.1 Mood Induction

For mood induction, one audio clip for high arousal and one for low arousal were selected. The two audio clips, Mozart Sonata for two pianos K448 (1985, track 1) for high arousal and Adagio by Albinoni [44] for low arousal, used in this research were validated for their arousal induction nature by Thompson et al., [45]. Thompson et al.

[45] observed that performance of the participants who listened to the music by Mozart (high arousal music) was better in spatial tasks compared to silent condition. They further observed that both piece of music induced different responses on enjoyment, arousal and mood measures. As was done in a study by Silvia and Abele [46], a neutral audio clip "Hymn" [47] was selected to induce neutral mood. In addition participants were instructed to bring their own high and low arousal audio clips, which they thought would put them in a low or high arousing state. Participants were not asked to bring neutral music.

4.1.2 Participants

Experiment took place in each individual's home with a quiet room with no interruptions. The aim of this setup was to avoid anxiety of coming to a lab. In addition it was assumed that they would be more relaxed in their own home environment. A total of 16 participants took part in this study. All the participants were male computer users with a mean age of 26 years (*SD* = 3.1). The participants had an average computer use experience of 7.3 years (*SD* = 2.8). Among participants 38% rated themselves as medium computer users, 31% rated themselves as expert computer users, and 31% rated themselves as programmers.

4.2 Study Setup

Department ethics committee approved this experiment. Consent was obtained from participants before conducting the experiment. Participants were asked to complete some programming tasks in Alice. Alice is an object-oriented language used to teach programming to children and was developed by Carnegie Mellon University. It is free to use and has some interesting aspects like visual object and visual construct and structures. All participants completed four learning tutorials present in the Alice Interface. The tutorials were about using Alice and its various functionalities as well as an introduction about simulating stories. Two Aesop stories ('The Hare and the Tortoise' and 'The Cow and the Frog') were assigned as tasks to each participant.

The study was composed of two sessions of 30-minute with a 10-minute break between them. Each session was further composed of three sub-sessions. One sub-session was of neutral music, and two sub-sessions of either high arousal or low arousal music. High or low arousal music sub-sessions were counterbalanced. Similarly personal music and music from the literature were also counterbalanced. Neutral music sub-session started with five minutes of no music at all followed by five minutes of neutral music. This was followed by either high arousal (brought by participants or taken from the literature in sequence one after another) or low arousal (brought by participants or taken from the literature in sequence one after another) music. There was a break of 10 minutes or more between two main sessions. The sequences of sessions as well as music sequences are illustrated in Table 5.

The music was played in the background on the participants' headphone so that they were able to listen to the music while completing the tasks. The use of a headphone and a separate room made sure no external interruptions affected the participants.

4.3 Results

4.3.1 Data Preparation

The experiment was divided into two sub-sessions separated by a gap of 10 minutes or more. The whole experiment took nearly an hour and ten minutes. GSR were sampled every 10 seconds in both of the sessions, thus recording about 180 responses in each session of 30 minutes with an average of 60 responses in each of the neutral, personal and literature music sessions. Similarly the second session also had 180 responses forming 360 responses from a single participant. At some points in the experiment, participants of course also had no interaction with the computer for more than 10 seconds in order to think or to write ideas down on paper. Exploring the log showed that only in one occasion did a participant have no interaction with the computer for more than one minute. In all other cases this never exceeded a minute. Therefore, it was decided to take the average of GSR responses in a minute and take the average of behavioural data recorded in that minute. This resulted in 30 responses from each sub-session forming 60 responses for the experiment.

An application was developed to count and analyse all the behavioural data within the target minute of GSR response. The behavioural variables considered for analysis were (1) number of events, (2) time between events, (3) total number of windows switched, (4) standard deviation of time between events, (5) total alphabetic and numeric keys pressed, (6) total mouse clicks, (7) total backspace/delete keys pressed and (8) total of all other keys pressed within that time. Each GSR measurement was then linked to the average of each behavioural measure recorded within that minute of GSR measurement. This resulted into 60 GSR responses each with their associated eight behavioural measurement values.

The purpose of GSR meter was to record electrical conductivity of the skin, which changes according to changes in emotional reactions. However when initially connected with the fingers, the reading started from minimum and gradually increased. Similarly a rapid and fast movement of hand also caused an increase or decrease in measurements. To reduce this effect the first step carried out on the data was to smooth it. This was done using a macro function called "Smooth"[6]. This function performs all standard smoothing methods of exploratory data analysis explained by [48] with a high degree of flexibility.

A command passed to the smoothing function to smooth was "3RSSH"[7]. Smoothing is included in the Exploratory Data Analysis (EDA) and is used in various studies of psychophysiological nature to reduce the noise (e.g. [49]).

Table 5: The four-session/music sequencing permutation participants were allocated to

Session 1 (30 min)					Session 2 (30 min)			
Sub-Session 1 (10 min)		Sub-Session 2 (10 min)	Sub-Session 3 (10 min)		Sub-Session 1 (10 min)		Sub-Session 2 (10 min)	Sub-Session 3 (10 min)
NoM (5min)	NeM (5min)	HAP	HAL	Break (10 minute) break	NoM (5min)	NeM (5min)	LAP	LAL
NoM (5min)	NeM (5min)	LAP	LAL		NoM (5min)	NeM (5min)	HAP	HAL
NoM (5min)	NeM (5min)	HAL	HAP		NoM (5min)	NeM (5min)	LAL	LAP
NoM (5min)	NeM (5min)	LAL	LAP		NoM (5min)	NeM (5min)	HAL	HAP

Note: NoM = No Music, NeM = Neutral Music, HAP = High Arousal Brought by person, LAP = Low arousal brought by person, HAL, High arousal indicated in literature, LAL = Low arousal indicated in literature

4.3.2 Analysis

The first part of the analysis was to test whether the music chosen from the literature or brought by the participants' had an effect on participants' GSR. Each participant's average GSR score in high arousal or low arousal session was calculated. A relative score was calculated by subtracting GSR measurements obtained when

[6] http://www.quantdec.com/Excel/smoothing.htm

[7] When a median smooth is immediately followed by an "R" (repeat) command, then continue to apply the median smooth until no more changes occur "Split" the sequence.
S = "Smooth": This process dissects the sequence into shorter subsequences at all places where two successive values are identical, applies a 3R smooth to each subsequence, reassembles them, and polishes the result with a 3 smooth.
H = "Hann" the sequence: This is convolution with a symmetrical kernel having weights (1/4, 1/2, 1/4). That is, each value x[i] is replaced by x[i-1]/4 + x[i]/2 + x[i+1]/4. The end values are not changed.

participants were listening to high arousal or low arousal music from the GSR measurements when participants were listening to neutral music or no music at all. The equation for this can be represented as follows:

$$GSR_{relative} = GSR_{neutral} - GSR_{low\ or\ high\ arousal}$$

A repeated measure Multivariate Analysis of Variance (MANOVA) was conducted to assess if there was a difference between high arousal and low arousal music, and music selected from the literature or brought by the participants on their GSR. No significant main effect ($F(1, 16) = 2.30$, $p = 0.15$) was found for high arousal versus low arousal. This means that together these two types of music did not induced uniformly either high or low arousal. However a significant effect ($F(1, 16) = 13.43$, $p = 0.002$) was found for the music type. Examining Fig 2, it seems that on average the music from the literature induced an arousal level close to the arousal level obtained in the neutral sessions (high arousal with a relative mean of 11.3 µS and a low arousal with a relative mean of -4.75 µS). The music brought by the participants resulted in much more high arousal (high arousal with a relative mean of -78.4 µS and low arousal with a relative mean of -24.4 µS). Also the analysis found a significant interaction effect ($F(1, 16) = 13.99$, $p = 0.002$) between music type and arousal level. It seems only the music brought by the participants significantly induced the expected relative levels of arousal. To confirm this, two paired sample *t*-tests were conducted as a post analysis. One pair contained readings of GSR while high arousal and low arousal was being induced via music brought by participants.

The second pair contained reading of GSR while high arousal and low arousal was being induced using selected music from literature. The results confirmed that the music brought by the participants significantly induced expected relative level of arousal ($t(16) = -4.9$, $p < 0.001$). However the effect of music selected from literature did not have significant effect ($t(16) = 0.8$, $p = 0.42$). The following analyses therefore only include the 20 samples obtained from the sessions using music brought by the participants.

To get an insight into whether a uniform generic mood measure across the participants would be possible, a within subject repeated MANOVA was conducted with arousal as independent variable with two levels (High vs. Low) for music brought by the participants. The dependent measures were the eight behavioural measure derived from keyboard and mouse data.

Table 6: Weighted correlations, Fisher's *z* scores and *p.* values of GSR and behavioural variables.

Measures	GSR							
	Events	Average Time Between Events	Windows Switched	SD Time b/w Events	Back space, delete Keys	Alphabetic numeric keys	Mouse Clicks	Others keys
Weighted r	-0.021	0.157	-0.013	-0.051	0.047	-0.066	0.057	0.023
Fisher z score	-0.022	0.176	-0.011	0.001	0.052	-0.070	0.063	0.026
p value	0.544	0.185	0.522	0.498	0.396	0.640	0.374	0.448

The analysis found no significant main ($F(8,8) = 1.03$, $p = 0.48$) effect for arousal on the behavioural keyboard and mouse data. The Univariate analyses also did not show a significant effect of arousal on the separate behavioural measures. This suggests that arousal difference did not show a clear uniform effect on keyboard and mouse behaviour across the participants in these two conditions.

To confirm above finding, eight individual Pearson correlations between 180 interaction behaviour samples and GSR samples were calculated. Next, weighted correlations and related Fisher's *z* score transformations were calculated. Table 6 shows these values and respective *p.* values. As in study 1, none of the correlations reached a significant level ($p > 0.05$). This again indicates that a generic mood measure from interaction behaviour of keyboard and mouse is unlikely.

As in study 1 the second step was to examine the possibility of individualized mood measures. For each participant a multiple regression analysis was conducted with as dependent variable GSR and as independent variables the eight interaction measures by using the Enter methods. Table 7 shows the individual correlations, the Fisher's *z* scores and their weighted averages. With a weighted correlation of 0.63, i.e. a large effect [38], and Fisher's *z* score of 0.75 the analysis again shows a significant ($Z = 3.34$, one-tailed $p. < 0.001$) relation between actual GSR values and the predicted GSR values based on keyboard and mouse data.

4.4 Discussion

The result seems to confirm the findings of the first study. No support was again found for a generic relationship between mood and interaction behaviour and but again results supported individual relationships. Still as in any empirical study, study 2 also had a number of limitations to reflect on. For example, there was a large variability in GSR, which is not uncommon as discussed by Gaillard and Kramer [50]. Data smoothing algorithms were used to remove this variability but it could not be removed completely. As [50] also pointed out, psychophysiological measures might also be affected by other factors than arousal like the workload and the strain of the participants' task.

However, by systematically assigning participants to the different music conditions it is unlikely that these factors would have caused the effects found. Another noise factor might have been the movement of the hands while typing or moving the mouse. Giving participants instructions to limit their hand movement would have interfered with natural interaction behaviour, which would have been undesirable for this study.

5 Conclusions, Discussions and Future Research

Moods and emotions are important factors in effective human computer interaction and so it is important to consider them and how they affect performance [51]. The results from both studies suggest that it is possible to measure individual moods based on their keyboard and mouse interaction. On the other hand, the conclusion of these two studies is also that creating a generic mood predictor based on the same data seems not possible at present day. Still, further research might prove otherwise. Therefore the main scientific contribution of the paper is the new empirical established insights, especially that an individualized mood predictor based on keyboard and mouse behaviour is possible. The two main advantages of this approach are: 1) unobtrusiveness, i.e. users do not have to wear special measuring equipment, or are not interrupted to complete a questionnaire; and 2) accessibility, i.e. no need to purchase additional equipment. The practical applications of such a relatively inexpensive and non-interruptive individual mood measure could be sought in situations where individuals require more insight in their own mood, or where computers could benefit from information about the user's mood in order to act on this. For example, people with bipolar disorder sometime use mood charts to identify patterns in their mood swings in an attempt to manage their illness. Potentially this subjective self-tracking could be supported with automated mood analysis of their computer interaction. Nowadays computer applications are more often designed to act as a social actor also, e.g. electronic coach, trainer, or assistant and therefore might also benefit from information about the mood state of the user. This would for example allow the computer to tailor its affective feedback to the mood of its user and thereby be more effective or persuasive [52]

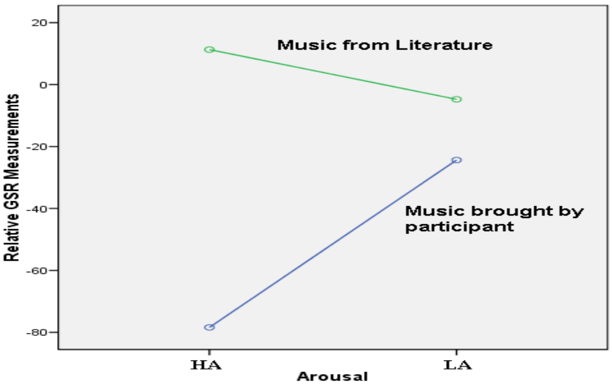

Fig 2: Relative Galvanic Skin Response (GSRrelative = GSRneutral – GSRlow/high) in low (LA) and high (HA) arousal conditions. Upper line represents music from the literature whereas line near the bottom represents music brought by person.

Table 7: Correlations and effect size of regression analyses on individual sample set to predict galvanic skin response

Participant ID	Individual sample size	*R*	Effect size
1	20	.55	0.62
2	20	.19	0.19
3	20	.79	1.08
4	20	.70	0.87
5	20	.56	0.63
6	20	.78	1.04
7	20	.74	0.95
8	20	.77	1.02
9	20	.63	0.73
10	20	.42	0.45
11	20	.62	0.73
12	20	.63	0.75
13	20	.52	0.57
14	20	.59	0.67
15	20	.50	0.54
16	20	.80	1.10
n meta-analysis		*weighted r*	*Fisher's z*
16		0.63	0.75

Like speech recognition technology, this individualized measure first needs to be trained to adapt to the individual user. The computer will learn by tracking keyboard and mouse behaviour and asking for subjective mood rating. Based on this, it could build an individualized model over time, which it then can utilize to predict the user's mood.

Besides multiple regression models, future research might explore whether other techniques would improve mood prediction; for example, using neural networks and other learning algorithms. This might increase the percentage of affective state classification. Zhai and Barreto [53] considered such work. They extracted some features from Skin Responses (GSR), Pupil Diameter (PD), and Blood Pressure Volume (BPV) and fed these features into three learning algorithms: Naïve Bayes Classifier, Decision Tree Classifier and Support Vector Machine (SVM). They found an accuracy of 78.65%, 88% and 90.1% respectively in classifying affective states. Future research could take this a step further in not only classifying the states but also classifying rating of valence and arousal; something the results of this study suggest is possible on an individual level. Another possibility of future work is to combine various interactional data like speech, images and video with keyboard interaction. Combined measures might increase estimation chance as research such as Pantic & Rothkrantz [54] and Chitu et. al. [55] demonstrated.

There are also some potential ethical issues that need to be considered. For example, users sometimes might not want a computer to know their mood and act on it. Also there are risks that other people besides the user can get this mood information, such as a colleague or an employer. Similarly an on-line shopping application that could detect someone's mood could use this to make persuasive sales offers. These issues might be addressed by giving users control over their mood measurement i.e. allowing user to enable and disable the mood measurement. While these are important issues that need to be addressed, the work presented in this paper suggests that individualized interaction based mood measurement might be possible, laying the foundation for a less intrusive mood measure instrument.

6 References

1. Brave S, Nass C. Emotion in Human-Computer Interaction: In The Human-Computer Interaction handbook: fundamentals, evolving technologies and emerging application, Eds. Jacko J.A, Sears A. Routledge Publishers, 2003.

2. Plutchik R. Emotion: A psychoevolutionary synthesis. New York: Harper and Row, 1980.

3. Picard R.W. Affective Computing, MIT Paper back edition 2000.

4. Zimmermann P, Guttormsen S, Danuser B. Gomez P. Affective Computing - A Rationale for Measuring Mood with Mouse and Keyboard. International Journal of Occupational Safety and Ergonomics. 2003, 9, 539-551

5. Klein J. Computer Response to User Frustration. Thesis, MIT Media Lab Technical Report, 1999, No. 480

6. Lazar J, Jones A, Hackley M, Shneiderman B. Severity and impact of computer user frustration: A comarison of student and workplace users. Interacting with Computers, 2006, 18(2), 187-207

7. Ross J. M, Zhang H. Structured Programmers Learning Object-Oriented Programming: Cognitive Considerations, SIGCHI Bulletin, 1997, 29(4)

8. Khan I.A, Brinkman W.-P, Hierons, R.M. Do moods affect programmers debug performance?. Cognition Technology & Work, 2010, 13(4), 245-258

9. Lang P. J. The Emotion Probe: Studies of Motivation and Attention. American Psychologist, 1995, 505, 372-385.

10. Buchanan T, Johnson J. A, Goldberg L.R. Implementing a five-factor personality inventory for use on the internet. European Journal of Psychological Assessment, 2005, 21, 115-127.

11. Valstar M, Patras I, Pantic M. Facial action unit recognition using temporal templates. 13th IEEE International Workshop on Robot and Human Interactive Communication, 2004.

12. Gendolla G. On the Impact of Mood on Behaviour: An Integrative Theory and a Review, Review of General Psychology, 2000, Vol. 4, pp. 378–408.

13. Forgas J.P, Fiedler K. Us and Them: Mood Effects on Intergroup Discrimination. Journal of Personality and Social Psychology, 1996, 70, 28-40.

14. Manucia G.K, Donald J.B, Robert B.C. Mood Influences on Helping: Direct Effects or Side Effects? Journal of Personality and Social Psychology, 1984, 46, 357–64.

15. V. C. Ottati and L. M. Isbell, "Effects on mood during exposure to target information on subsequently reported judgments: An on-line model of misattribution and correction," Journal of Personality and Social Psychology, vol. 71, p. 39, 1996.

16. Isen A. M, Stanley F. S. The Effect of Feeling Good on a Helping Task That Is Incompatible with Good Mood. Social Psychology, 1978,41, 346-349.

17. Isen A. M. Nehemia G. The Influence of Positive Affect on Acceptable Level of Risk: The Person with a Large Canoe Has a Large Worry, Organizational Behaviour and Human Decision Processes, 1987, 39, 145–54.

18. Nygren T.E, Alice M. I, Pamela J.T, Jessica D. The Influence of Positive Affect on the Decision Rule in Risk Situations: Focus on Outcome and Especially Avoidance of Loss Rather than Probability, Organizational Behaviour and Human Decision Processes, 1996, 66, 59–72.

19. Pham M.T. Representativeness, Relevance, and the Use of Feelings in Decision Making, Journal of Consumer Research, 1998. 25, 144–159.

20. Johnson N, Galata A, Hogg D. The acquisition and use of interaction behaviour models. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1998, 866–871.

21. Gavrila D. M. Davis, L. S. Towards 3-D model-based tracking and recognition of human movement: a multi-view approach. In Proc. International Workshop on Automatic Face and Gesture Recognition, 1995, 272–277.

22. Lefter I, Rothkrantz L. J. M, , Wiggers P, Van Leeuwen D. A. Emotion Recognition from Speech by Combining Databases and Fusion of Classifiers. Text, Speech and Dialogue. Lecture Notes in Artificial Intelligence, 2010, 6231, 353-360

23. Bobick A, Davis J.W. Action recognition using temporal templates. In CVPR, 1997, 125–146.

24. Sakurazawa S, Yoshida N, Munekata N, Omi A, Takeshima H, Koto H, Gentsu K, Kimura K, Kawamura K, Miyamoto M, Arima R, Mori T, Sekiya T, Furukawa T, Hashimoto T, Numata H, Akita J, Tsukahara Y, Matsubara H. A Computer gaming using galvanic skin response. ACM International Conference Proceeding Series, 2003. 38, 1-3.

25. Backs R.W, Boucsein W. Engineering Psychophysiology as a Discipline: Historical and Theoretical Aspects in Engineering Psychophysiology. Eds. Backs R.W, Boucsein W. 1st Ed. Lawrence Erlbaum Publications, London, 2000, 3-30.

26. Chanel G, Ansari-Asl K, Pun T. Valence-arousal evaluation using physiological signals in an emotion recall paradigm, IEEE International Conference on Systems, Man & Cybernetics, 2007, 2662-2667.

27. Boucsein W, Thum M. Design of work\rest schedules for computer work based on psychophysiological recovery measure. International Journal of Industrial Ergonomics, 1997, 20, 51-57.

28. Haider E, Luczak H, Rohmert W. Ergonomics investigations of workplaces in a police command-control centre equipped with TV displays. Applied Ergonomics, 1982, 13, 163-170.

29. Schleife L.M, Ley R. End tidal PCO as an index of psychophysiological activity during VDT data-entry work and relaxation. Ergonomics, 1994, 37, 245-254

30. Lissetti C.L, Nasoz F. MAUI: a Multimodal affective user interface. Proceedings of the tenth ACM international conference on Multimedia, Juan-les-Pins, France, 2002, 161-170.

31. Mahr W, Carlsson R, Fredriksson J, Maul O, Fjeld M. Tabletop Interaction. Research Alert. Proc. ACM NordiCHI 2006, 499-500.

32. Khanna P, Sasikumar M. Recognising emotions from keyboard stroke Patterns. International journal of computer applications, 2010, 119, 1-5.

33. Tsihrintzis G.A, Virvou M, Alepis E, Stathopoulou I.O. Towards Improving Visual-Facial Emotion Recognition through Use of Complementary Keyboard-Stroke Pattern Information. In: International Conference on Information Technology: New Generations (ITNG), 2008, 32-37, doi: 10.1109/ITNG.2008.152

34. Russell J.A. A Circumplex Model of Affect. Journal of Personality and Social Psychology, 39(6): pp. 1161-1178.

35. Bradley M.M, Lang, P.J. Measuring emotion: The self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry, 1994, 25, 49-59.

36. Mehrabian A. Basic Dimensions for a General Psychological Theory: OG&H Publishers. 1980.

37. Lang P. J. Behavioural treatment and bio-behavioural assessment: computer applications, In: Sidowski J.B, Johnson J.H, Williams T.A. Eds. Technology in mental health care delivery systems. Norwood, NJ: Ablex, 1980.

38. Cohen J. Statistical Power Analysis, Journal of Clinical Psychiatry, 1992, 1(3), 61-69.

39. Card K.C, Moran T.P. Newell A. The Psychology of Human-Computer Interaction. Lawrence Erlbaum Publishers, Hillsdale, New Jersey, 1983.

40. Borenstein M, Hedges L.V, Higgins, J.P.T, Rothstein H.R. Introduction to meta-analysis. Chichester, UK: Wiley, 2009.

41. Wang H, Prendinger H, Igarashi T. Communicating emotions in online chat using physiological sensors animated text. In: Conference on Human Factors in Computing Systems CHI, 04 Vienna Austria, 2004, 1171-1174

42. Remington N.A, Fabrigar L.R. Re-examining the Circumplex model of affect. Journal of Personality and Social Psychology, 2000, 79(2), 286-300.

43. Shepherd P. Tools for Transformation, trans4mind.com, URL:=http://www.trans4mind.com/metercourse/GSR_1.html, 2001, accessed at 12/10/2010

44. Albinoni T.G. Adagio in G Minor for organ and strings [Recorded by SolistiVeniti, Conducted by Scimone,

C.], 1981, On Albinoni's Adagios [CD]. Perivale, England Warner Classics, 1996

45. Thompson W.F. Schellenberg E.G. Husain G. Arousal, Mood and the Mozart effect, Psychological Science, 2001, 12(3).

46. Silvia P.J, Abele A.E. Can positive affect induce self-focused attention? Methodological and measurement issues. Cognition and Emotion, 2002, 16(6), 845-853.

47. Moby. Hymn. On Everything is wrong [CD]. New York: Elektra, 1995

48. Tukey J. Exploratory Data Analysis, Addison-Wesley Press, 1977.

49. Baumgartner R, Ryner L, Somorjai R. Summers R. Exploratory Data Analysis Reveals Spatio-temporal Structure of Null fMRI Data, Proc. Intl. Sot. Mag. Reson. Med., 2000, 8.

50. Gaillard A. W. K, Kramer A. F. Theoretical and methodological issues in psychophysiological research. In Backs R. W, Boucsein W. Eds, Engineering psychophysiology, 2001, 31–58 Mahwah, NJ: Erlbaum.

51. Harper R, Rodden T, Rogers Y. Sellen A. Being Human: Human Computer Interaction in 2020, HCI 2020, Microsoft Research meeting El Bulli Hacienda Hotel, Sveille, Spain., 2007.

52. Jenniger R. Evaluating the consequences of feedback in intelligent tutoring systems. Third International Conference on Affective Computing and Intelligent Interaction and Workshops ACII, North Carolina State Univ., Raleigh, NC, USA. 2009.

53. Zhai J, Barreto A. Stress Detection in Computer Users through Non-Invasive Monitoring of Physiological Signals, Biomedical Science Instrumentation, 2006, 42, 495-500.

54. Pantic M, Rothkrantz L.J.M. Toward an Affect-sensitive Multimodal Human-Computer Interaction, Proceedings of the IEEE, 2003, 91(9), 1370-1390.

55. Chitu A.G, Rothkrantz L.J.M, Wojdel J.C, Wiggers P. Comparison between Different Feature Extraction Techniques for Audio-Visual Speech Recognition. Journal on Multimodal User Interfaces, 2007, 1(1), 7-20