Assessing NeuroSky's Usability to Detect Attention Levels in an Assessment Exercise

Genaro Rebolledo-Mendez[1,3], Ian Dunwell[1], Erika A. Martínez-Mirón[2], María Dolores Vargas-Cerdán[3], Sara de Freitas[1], Fotis Liarokapis[4], and Alma R. García-Gaona[3]

[1] Serious Games Institute, Coventry University, UK
[2] CCADET, UNAM, Mexico
[3] Facultad de Estadistica e Informatica, Universidad Veracruzana, Mexico
[4] Interactive Worlds Applied Research Group, Coventry University, UK
{GRebolledoMendez,IDunwell,Sfreitas, F.Liarokapis}@cad.coventry.ac.uk,
erika.martinez@ccadet.unam.mx, {dvargas,agarcia}@uv.mx

Abstract. This paper presents the results of a usability evaluation of the NeuroSky's MindSet (MS). Until recently most Brain Computer Interfaces (BCI) have been designed for clinical and research purposes partly due to their size and complexity. However, a new generation of consumer-oriented BCI has appeared for the video game industry. The MS, a headset with a single electrode, is based on electro-encephalogram readings (EEG) capturing faint electrical signals generated by neural activity. The electrical signal across the electrode is measured to determine levels of attention (based on Alpha waveforms) and then translated into binary data. This paper presents the results of an evaluation to assess the usability of the MS by defining a model of attention to fuse attention signals with user-generated data in a Second Life assessment exercise. The results of this evaluation suggest that the MS provides accurate readings regarding attention, since there is a positive correlation between measured and self-reported attention levels. The results also suggest there are some usability and technical problems with its operation. Future research is presented consisting of the definition a standardized reading methodology and an algorithm to level out the natural fluctuation of users' attention levels if they are to be used as inputs.

1 Introduction

This paper presents a usability evaluation of NeuroSky's MindSet (MS) device. An aspect of interest was to investigate whether MS readings can be combined with user-generated data. The amalgamation of physiological and user-generated data would allow the programming of more sophisticated user models. An experimental setting was set up to analyze MS usability in an assessment exercise in Second Life. The assessment [10] is based on a multiple-choice questionnaire in the area of programming for Computer Science undergraduate students. The questionnaire is presented by an Artificial Intelligence-controlled avatar (AI-avatar) who is aware of the levels of attention of the

J.A. Jacko (Ed.): Human-Computer Interaction, Part I, HCII 2009, LNCS 5610, pp. 149–158, 2009.

person interacting with it. The MS[1] also provides a measurement of the user's meditative state (derived from alpha wave activity). In this paper, however, only the levels of attention are used, given their role and importance in educational settings. The objective of this study is threefold: firstly, the MS general usability is examined. Secondly, an analysis of how well it is possible to fuse information generated as part of normal interactions with brain activity. Thirdly, an analysis of the MS adaptability to different able-users is provided. The significance of this work lies in that it presents evidence of the usability of a commercially available BCI and its suitability to be incorporated into serious games. The paper is organized in five sections. Section two presents a literature review about Brain Computer Interfaces and their use for learning. Section three describes the Assessment exercise used as test bed and presents the materials, participants and methodology followed during the evaluation. Section four presents the results of the evaluation and, finally, section five provides the conclusions and future research.

2 Brain Computer Interfaces (BCI)

Brain Computer Interface (BCI) technology represents a rapidly emerging field of research with applications ranging from prosthetics and control systems [6] through to medical diagnostics. This study only considers BCI technologies that use sensors that measure and interpret brain activity (commonly termed neural bio-recorders [14]) as a source of input. The longest established method of neural bio-recording, developed in 1927 by Berger [3], is the application of electrodes that measure the changes in field potential over time arising from synaptic currents. This forms the basis for EEG. In the last two decades, advances in medical imaging technology have presented a variety of alternative means for bio-recording, such as functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG), and positron emission tomography (PET). A fundamental difference between bio-recording technologies used for diagnostic imaging, and those used for BCI applications, is a typical requirement for real or quasi-real time performance in order to translate user input into interactive responses. In 2003, a taxonomy by Mason and Birch [8] identified MEG, PET and fMRI as unsuitable for BCI applications, due to the equipment required to perform and analyze the scan in real-time, but more recent attempts to use fMRI as a BCI input device have demonstrated significant future potential in this area [12].

Bio-recording BCIs have become a topic of research interest both as a means for obtaining user input, and studying responses to stimuli. Several studies have already demonstrated the ability of an EEG-based BCI to control a simple pointing device similar to a mouse [9, 12] and advancing these systems to allow users more accurate and responsive control systems is a significant area for research. Of particular interest to this study is the use of BCI technologies in learning-related applications. The recent use of fMRI to decode mental [4] and cognitive [11] states illustrates a definite capability to measure affect through bio-recording, but the intrusiveness of the scanning equipment makes it difficult to utilize the information gained to provide feedback to a user performing typical real-world learning activities.

In this study, the effectiveness of one of the first commercially available lightweight EEG devices, NeuroSky's MS, is considered. Via the application of a single

[1] The MB is a developer-only headset. NeuroSky's newest headset has been designed to address comfort and fitting problems and is available to both developers and consumers.

electrode and signal-processing unit in a headband arrangement, the MS provides two 100-state outputs operating at 1Hz. These outputs are described by the developers as providing separate measures of 'attention' and 'meditation', and it is thus assumed these readings are inferred from processing beta and alpha wave activity respectively. Although the MS provides a much coarser picture of brain activity than multi-electrode EEG or the other aforementioned technologies, the principle advantage of the MS is its unobtrusive nature, which minimises the aforementioned difficulties in conducting accurate user studies due to the stress or distraction induced by the scanning process. Research into EEG biofeedback as a tool to aid individuals with learning difficulties [5] represents an area for ongoing study, and the future widespread availability of devices similar to the MS to home users presents an interesting opportunity to utilize these technologies in broader applications.

3 An Assessment Exercise in Second Life

An assessment exercise was developed to examine the MS. The exercise works in combination with a model of attention [10] built around dynamic variables generated by the learner's brain (MS inputs) and the learner's actions in a computer-based learning situation. The combination of physiological (attention) and data variables is not new [7, 1]. Our approach, however, fuses MS readings (providing a more accurate reading of the learner's attention based on neural activity) with user-generated data. In our model, attention readings are combined with information such as the number of questions answered correctly (or incorrectly), or the time taken to answer each question, to model attention within the assessment exercise.

The MS reads attention levels in an arbitrary scale ranging from 0 to 100. There is an initial delay of between 7 and 10 seconds before the first value reaches the computer and newer values of attention are calculated at a rate of 1Hz (one value per second, see Figure 1). A value of -3 indicates no signal is being read and values equal to or greater than 0 indicate increasing levels of attention with a maximum value of 100. Given the dynamic nature of the attention patterns and the potentially large data sets obtained, the model of attention underpinning the assessment exercise is associated to a particular learning episode lasting more than one second. The model of attention not only determines (detects) attention patterns but also provides (reacts) feedback to the learner [10].

The assessment exercise consists of presenting a Second Life[2], AI-driven avatar able to pose questions, use a pre-defined set of reactions and have limited conversations with learners in Second Life. The AI-driven avatar was programmed using C# (C-sharp) in combination with the lib second life library[3]. Lib Second Life is a project aimed at understanding and extending Second Life's client to allow the programming of features using the C# programming languages. This tool enables the manipulation of avatars' behaviors so that they respond to other avatars'. To do so, the AI-driven avatar collects user-generated data during the interaction including MS inputs. The current implementation of the AI-driven avatar asks questions in a multiple-choice

[2] http://secondlife.com/

[3] http://www.libsecondlife.org/

format, while dynamically collecting information (answers to questions, time taken to respond, and whether users fail to answer). The data generated by the MS is transmitted to the computer via a USB interface and organized via a C# class which communicates with the AI-driven avatar. In this way, the model of attention is updated dynamically and considers input from the MS as well as the learner's performance behavior while underpinning the AI-driven avatar's behavior.

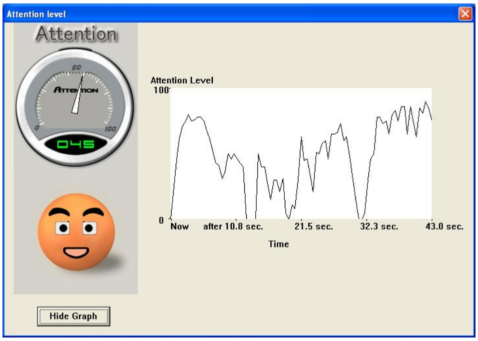

Fig. 1. Attention readings as read by the NeuroSky

For the purposes of assessing MS's usability, the assessment exercise consisted of ten questions in the area of Informatics, specifically for the area of Algorithms. This area was targeted since it has been noted first year students in the Informatics department often struggle with the conception and definition of algorithms, a fundamental part of programming. The assessment exercise asked nine theoretical questions and presented three possible answers. For example, the avatar would ask 'How do you call a finite and ordered number of steps to solve a computational problem?' while offering 'a) Program, b) Algorithm, c) Programming language' as possible answers. The assessment exercise also includes the resolution of one practical problem, answered by the learner by hand while still wearing the MS.

3.1 Materials

To evaluate the MS's reliability, two adaptations of the Attention Deficit and Hyperactivity disorder (ADHD) test and a usability questionnaire were defined. The attention tests consisted of seven items based on the DSV[4]-IV criteria [2]. The items chosen for the attention test were: 1. Difficulty to stay in one position, 2. Difficulty in sustaining attention, 3. Difficulty to keep quiet often interrupting others, 4. Difficulty to follow through on instructions, 5. Difficulty to organize tasks and activities, 6. Difficulty or avoidance of tasks that require sustained mental effort and 7. Difficulty to listen to what is being said by others.

Each item was adapted to assess attention both in the class and at interaction time. To answer individual questions, participants were asked to choose the degree which they believed reflected their behavior in a Likert type scale with 5 options. For example, question 1 of the attention questionnaire asked the participant: 'How often is it difficult for me to remain seated in one position whilst working with algorithms in class/during the interaction?' with the answers 1) all the time, 2) most of the time, 3) some times, 4) occasionally and 5) never. Note that for both questionnaires the same seven questions were asked but were rephrased considering the class for the pre-test

[4] Diagnostic and Statistical Manual of Mental Disorders.

or the interaction for the post-test. The usability questionnaire consisted of adapting three principles of usability into three questions (a) comfort of the device; (b) easiness to wear; and (c) degree of frustration.

To answer the usability questionnaire participants were asked to select the degree to which they felt the MS faired during the interaction via a Likert type scale with 5 options. For example, question 1 of the usability questionnaire asked the student: 'Was using Neurosky' 1) Very uncomfortable, 2) Uncomfortable, 3) Neutral, 4) Comfortable, 5) Very comfortable. Note that to report the usability of the MS, other factors were also considered such as battery life, light indicators and data read/write times and intervals.

3.2 Participants and Methodology

An evaluation (N=40) to assess the usability of the MS was conducted among first-year undergraduate students in the Informatics Department at the University of Veracruz, Mexico. The population consisted of 28 males and 12 females, 38 undertaking the first year of their studies and 2 undertaking the third year. 26 students (65%) of the population were 18 years old, 12 students (30%) were 19 years old and 2 students (5%) were 20 years old. The participants interacted with the AI-avatar for an average of 9.48 minutes answering ten questions posed by the avatar within the assessment exercise (see previous section). During the experiment, the following procedure was followed: 1) students were asked to read the consent form, specifying the objectives of the study and prompted to either agree or disagree, 2) students were asked to solve an online pre-test consisting of the adaptation of the attention deficit and hyperactivity disorder (ADHD) questionnaire to assess their attention levels in class, 3) students were instructed on how to use the learning environment, and finally 4) the students were asked to answer an online post-test consisting of the usability questionnaire and the adaptation of the ADHD questionnaire to assess their attention levels *during* the interaction in the assessment exercise. Individual logs registering the students' answers and attention levels as read by the MS were kept for analyses. All students agreed to participate in the experiment but in some cases (N=6) the data was discarded since the MS did not produce readings for these participants. See the results section for a description of these problems. Cases with missing data were not considered in the analysis.

4 Results

The results of this evaluation are organized to consider the MS's usability, how well the model fuses user-generated data and attention readings and the MS's adaptability.

4.1 Usability and Appropriateness of MS for Assessment Exercises

The main aspect of interest was MS's usability considering the responses to three questions (see materials section). This questionnaire considered three aspects to assess the usability of new computer-based devices: Comfort, Ease of Use, and the Degree of Frustration. The answers to the questionnaire are organized around each aspect considered. There was one question associated to every usability aspect.

Comfort

The results showed that for 5% (N=2) the MS was uncomfortable, for 10% (N=4) somewhat uncomfortable, for 35% (N=14) neither comfortable nor uncomfortable, for 25% (N=10) somewhat comfortable and for 25% (N=10) comfortable.

Ease of Use

The results showed 15% (N=6) students found the MS difficult to wear, 12.5% (N=5) found it somewhat difficult to wear, 37.5% (N=15) thought it was neither easy nor difficult to wear, 12.5% (N=5) found it somewhat easy to wear and 22.5% (N=9) thought it was easy to wear.

Degree of Frustration

The answers showed 2.5% (N=1) found the experience frustrating, 2.5% (N=1) thought it was somewhat frustrating, 22.5% (N=9) found the experiment neither frustrating nor satisfactory, 25% (N=10) thought it was somewhat satisfactory and 47.5% (N=19) had a satisfactory experience using the MS.

There were three aspects that only became apparent once the evaluation was over. The first aspect of interest was in relation to the pace and the way readings were collected. The attention model [1] considered readings in the space of time used by learners to formulate an answer for each question. The pace in which data was collected by the model was 10Hz which produced repeated measurements in some logs. This method of collecting data is inefficient as plotting attention fluctuations considering fixed, regular intervals might be difficult. People interested in programming the MS device should consider that, due to a hardware processing delay, the MS outputs operate at 1Hz, and need to program their algorithms accordingly. The second aspect of interest is in relation to difficulties wearing the device. When connection is lost, there is a delay of 7-10 seconds before a new reading is provided. Designers should consider this as a constant input might not be possible. The third aspect refers to MS's suitability as an input device for interface control. Developers need to consider that attention levels (and associated patterns) vary considerably between users (see Figure 2), as expected. If developers employ higher levels of attention as triggers for interface or system changes, they should consider some users normally have higher levels of attention without being prompted to put more attention. This normal variability creates the need to research and develop an algorithm to level-out initial differences in attention levels and patterns. On a related topic, MS's readings vary in a scale from 0 to 100 (see Figure 1): however, it is not yet clear what relationship exists between wave activity and processed output, whether the scale is linear, or whether the granularity of the 100-point scale is appropriate for all users. Finally, there were some usability problems that caused data loss, in particular:

1. In 3 cases the MS did not fit the participant's head properly leading to adjustments by the participants leading to not constant and unreliable readings. Another problem was people with longer hair having problems wearing the device to allow sensors touch the skin behind the ears at all times. During the experiment, extra time was required to make sure people with longer hair placed the device adequately.
2. In other 3 cases the MS ran out of battery. The battery was checked before each participant interacted with the assessment exercise using NeuroSky software via its associated software. However, despite the precautions taken and after having checked the green light on one of the device's side, battery life was very short. The

device does not alert the user when battery levels are low, so it was not clear when batteries needed to be replaced. This was a problem at the beginning of the experiment but later on batteries were replaced on daily basis.

4.2 Adaptability to Different Users

One of the characteristics of the MS reader is that it can be worn by different users producing different outputs. This would allow for adaptation of the model [10] in the frame of the assessment exercise. It was expected MS outputs would vary for different users reflecting varying levels of attention. Furthermore, this adaptation would be fast and seamless without the need to train the device for a new user. To throw some light onto the issue of adaptability, it was speculated attention readings would be different for individuals. It was also hypothesized there would be a positive correlation between the readings and the self-assessment attention test (see materials section).

To assess variability among participants, a test of normality was done to see the distribution of the participants' average attention levels. Table 1 shows descriptive statistics of the readings for the population (N=34). The results of a test of normal distribution showed that the data is normally distributed (Shapiro-Wilk = .983, p = .852) suggesting there is not a tendency to replicate particular readings. Figure 2 illustrates the Q-Q plot for this sample suggesting a good distribution of average attention levels during the assessment exercise.

Table 1. Descriptive statistics

	N	Min	Max	Mean	Std. Dev.
Student's attention levels	34	14.99	88.00	53.40	16.69
Student's self reported attention	34	3.0	5.0	4.27	.44548

Another test designed to see whether MS readings adapted to individual participants, was a correlation between the readings and the self-reported attention using the post-test questionnaire. A positive correlation was expected between these two variables.

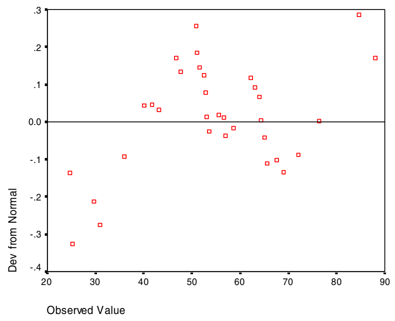

Fig. 2. Q-Q plot of students' average attention levels during the assessment exercise

Table 1 shows the descriptive statistics for the two variables. To calculate self-reported attention levels, the mean of the answers to the 7 items of the attention post-test was calculated per participant; lesser values indicate lesser attention levels. The results of a Pearson's correlation between the two variables indicated a significant, positive correlation (Pearson's = -.391, p = .022).

4.3 Fusing User-Generated Information with MS Readings

One way to analyze whether the data was fused correctly was to check the logs for missing or incorrect data. The results of this analysis showed that there were six participants (15%, original sample N=40) for which the MS did not produce accurate readings. An analysis of the logs for the remaining participants (N=34) showed the device produced readings throughout the length of the experiment (average time = 9.48 minutes) without having an erroneous datum (attention = -3). The causes for the lack of readings in 6 cases were due to usability problems (see following section).

Another way of throwing some light on how well the MS readings and user-generated data were fused consisted of analyzing the logs to see whether there was a variation on the model's reactions for the sample. Since the reactions given by the AI avatar could be of six types [1], the frequency was calculated for each reaction type for the entire population with correct NeuroSky readings (N=34), see Table 2.

Table 2. Frequencies associated to the model's reaction types for the population (N=34)

Reaction Type	6	5	4	3	2	1
Frequency	128	172	0	77	13	0

It was expected the frequencies for reaction types 4 and 1 would be 0 given the averages of the four binary inputs. Reaction Types 5 was the most common type followed by Reaction Types 6, 3 and 2. Given the 8 possible results of averaging out the four binary inputs [10], it was expected Reaction Type 3 would be the most frequent. However, this was not the case suggesting the model did vary and the reactions type provided were in accordance to the variations in attention, time, and whether answers were correct.

Finally, the responses to two questions in the post-test questionnaire gave an indication of students' subjective perceptions about how well the reaction types were adequate to their attention needs. The first question asked: 'how frequently the reactions helped you realize there was something wrong with the way you were answering the questions?' The answers showed 25% (N=10) of students felt the avatar helped them all the time, 20% (N=8) said most of the time, 35% (N=14) mentioned some times, 12.5% (N=5) said rarely and 7.5% (N=3) stated never. The second question asked 'how appropriate they thought the combined use of MS and avatars was appropriate for computer-based educational purposes?' Students' answered with 65% (N=26) saying it was appropriate, 12.5% (N=5) saying it was appropriate most of the time, 15% (N=6) saying it was neither appropriate nor inappropriate, 2.5% (N=1) saying it was somewhat inappropriate and 5% (N=2) saying it was inappropriate.

5 Conclusions and Future Work

The reliability of MS readings to assess attention levels and to amalgamate with user-generated data was evaluated in an assessment exercise in Second Life, N=34. The results showed there is variability in the readings and they correlate with self-reported attention levels suggesting the MS adapts to different users providing accurate readings of attention. The results of analyzing the device's usability suggest some users have problems with wearing the device due to head sizes or hair interference and that the device's signals to indicate flat batteries are poor. By analyzing individual logs it was possible to determine that, when the device fits properly, the MS provides valid and constant data as expected. Log analyses also helped establish the frequency different reactions types were provided in the exercise in the light of attention variability. The frequencies suggested the model did not lean to the most expected reaction (Type 3) but that it tended to be distributed amongst Reaction Types 5 and 6, providing an indication that user-generated data was fusing adequately with attention readings. When asked about their experience, 35% of the population said the avatar helped them realize there was something wrong with how s/he was answering the questions and 65% indicated using a MS in combination with avatars was appropriate in computer-based educational settings. When asked about comfort, 35% thought the device was neither comfortable not uncomfortable, 37.5% thought it was neither easy nor difficult to wear and 47.5% said they had a satisfactory experience with the device. There were other results that were apparent only after the evaluation. In particular, it was found that: 1) sampling rates need to be considered to organize data in fixed, regular intervals to determine attention. 2) Developers need to be aware there is a delay when readings are lost due to usability issues. 3) Variability imposes new challenges for developers who wish to use levels of attention as input to control or alter interfaces. Work for the future includes the combination of MS readings with other technologies such as using learner's gaze, body posture and facial expressions to read visual attention. Future work will be carried out to find out the degree of attention variability to program an algorithm capable of leveling-out different patterns of attention. In addition, future work will explore how attention data can be used to develop learner models that help understanding attention and engagement for informing game-based learning design and user modeling.

Acknowledgements

The research team thank the NeuroSky Corporation for providing the device for testing purposes. We also thank the students, lecturers and support staff at the Faculty of Informatics, Universidad Veracruzana, Mexico.

References

1. Amershi, S., Conati, C., McLaren, H.: Using Feature Selection and Unsupervised Clustering to Identify Affective Expressions in Educational Games. In: Workshop in Motivational and Affective Issues in ITS, 8th International Conference on Intelligent Tutoring Systems, Jhongli, Taiwan (2006)
2. Association, A.P.: Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Press (1994)
3. Berger, H.: On the electroencephalogram of man. In: Gloor, P. (ed.) The fourteen original reports on the human electroencephalogram, Amsterdam (1969)
4. Haynes, J.D., Rees, G.: Decoding mental states from brain activity in humans. Nature Neuroscience 7(7) (2006)

5. Linden, M., Habib, T., Radojevic, V.: A controlled study of the effects of EEG biofeedback on cognition and behavior of children with attention deficit disorder and learning disabilities. Applied Psychophysiology and Biofeedback 21(1) (1996)
6. Loudin, J.D., et al.: Optoelectronic retinal prosthesis: system design and performance. Journal of Neural Engineering 4, 72–84 (2007)
7. Manske, M., Conati, C.: Modelling Learning in an Educational Game. In: 12th Conference on Artificial Intelligence in Education, IOS Press, Amsterdam (2005)
8. Mason, S.G., Birch, G.E.: A general framework for brain-computer interface design. IEEE Transactions on Neural Systems and Rehabilitation Engineering 11, 70–85 (2003)
9. Poli, R., Cinel, C., Citi, L., Sepulveda, F.: Evolutionary brain computer interfaces. In: Giacobini, M. (ed.) EvoWorkshops 2007. LNCS, vol. 4448, pp. 301–310. Springer, Heidelberg (2007)
10. Rebolledo-Mendez, G., De Freitas, S.: Attention modeling using inputs from a Brain Computer Interface and user-generated data in Second Life. In: The Tenth International Conference on Multimodal Interfaces (ICMI 2008), Crete, Greece (2008)
11. Sona, D., Veeramachaneni, S., Olivetti, E., Avesani, P.: Inferring cognition from fMRI brain images. In: de Sá, J.M., Alexandre, L.A., Duch, W., Mandic, D.P. (eds.) ICANN 2007. LNCS, vol. 4669, pp. 869–878. Springer, Heidelberg (2007)
12. Sitaram, R., et al.: fMRI Brain-Computer Interfaces. IEEE Signal Processing Magazine 25(1), 95–106 (2008)
13. Trejo, L.J., Rosipal, R., Matthews, B.: Brain-computer interfaces for 1-D and 2-D cursor control: designs using volitional control of the EEG spectrum or steady-state visual evoked potentials. IEEE Transactions on Neural Systems and Rehabilitation Engineering 14(2), 225–229 (2006)
14. Vaughan, T., et al.: Brain-computer interface technology: a review of the second international meeting. IEEE Transactions on Neural Systems and Rehabilitation Engineering 11(2), 94–109 (2003)