A Physiological Approach for User Experience Evaluation

Case Study: Interaction with Tabletop Systems for Music Creation

Diana Margarita Mundó Spataro

MASTER THESIS UPF / 2012

THESIS DIRECTOR

Dr. Sergi Jordà (Music Technology Group – MTG)

THESIS SUPERVISOR

Sebastián Mealla (Music Technology Group – MTG)

DEPARTAMENTO DE TECNOLOGÍAS DE LA INFORMACIÓN Y LA COMUNICACIÓN (DTIC)

UNIVERSITAT POMPEU FABRA

Acknowledgment

First, I want to offer my sincere gratitude to my supervisor Sebastián Mealla for his support, dedication, advices and knowledge during the entire project. Furthermore, I want to thank Sergi Jordà, my thesis coordinator, for his helpful criticism and time dedicated to this project.

To Narcís Parés and Sylvain Le Groux for their time and comments during presentations.

To my parents, Diana Spataro and Juan Mundó, and my sister Maria Elvira Mundó, for their support and comprehension, and for making this master possible.

To my family in Barcelona, Quim, Jordi, Tabea, Karo, Jose Luis and Ana Mari, who did everything what was necessary for helping me during the entire process.

Finally, I want to thank to all my master's partners Nuch, Joseph, Aditya, Nicolás, Saavas, Ana, Lola, Merve, Ayse Naz, Sebastián, Jing, Penny, Matylda, Pili, Carlos, Diana, Marte, Sinan and Fabio for their support.

Abstract

Based on previous studies, physiological data has proven to be a promising, valuable indicator of emotional different states during game experience, such as emotional response and relaxation. However, there are few studies showing how such techniques can support and/or complement traditional User Experience (UX) assessment methods, both subjective (e.g. surveys and psychometric questionnaires) and objective (log files, tracking systems, etc.). We propose to develop a multimodal UX framework that combines both subjective, self-reported measures and objective, physiology based metrics, specifically EEG. We hypothesize that physiological measures can improve traditional UX evaluation methods by removing biases such as lack of memory of participants, social desirability, and adding real-time biometric feedback to asynchronous evaluation of interactive experience (e.g. post test questionnaires). In order to test the strength and possible applications of such multimodal assessment framework, we have designed a set of pilot tests and task-oriented experiments using tabletop tangible interfaces for music creation, where EEG data from subjects was recorded. Through this approach we aim at investigating whether significant correlations and interactions occur between subjective and physiological measures describing user's responses to this interactive music composition experiences in terms of challenge, enjoyment and excitement.

Keywords

User Experience (UX) evaluation, EEG, Questionnaires, TUIs, BCI, physiology, music, tabletops

Index

1. Introduction

Physiological signals, particularly encephalography (EEG) and heart rate (HR), have shown to be correlated with different psychological states such as mood states (Wingrave et al, 2011), emotional states - frustration, stress and relaxation (Nijholt et al, 2009; Pope & Stevens, 2011) - and motor imagery (Tangerman et al, 2009; Pfurtscheller et al, 2010). However, there are few studies in the field of Human-Computer Interaction (HCI) showing how such techniques support and/or complement traditional experience assessment methods, both subjective (eg. surveys and psychometric questionnaires) and objective (eg. log files and motion tracking systems). Moreover, most of these studies are focused on Game Experience evaluation (Nacke et al, 2010; Kivikangas et al, 2010) rather than general UX evaluation.

For this project we will use a multimodal framework of evaluation that includes as physiological measures EEG and as subjective, self-reported measures, the variable of Challenge from the Game Experience Questionnaire (GEQ) (IJsselsteijn et al) and the variables of Enjoyment and Excitement based on Giakoumis's questionnaire (2009). Using biometric signals correlated to subjective, self-reported data to evaluate UX when interacting with a tabletop, might produce more accurate results about performance and subjective experience by reducing biases directly related to subjective measures (eg. lack of memory of the participant or social desirability).

2. State of the Art

In this section we will describe both subjective and objective methods used in User Experience (UX) evaluation.

2.1 User Experience

User Experience (UX) has been a matter of research on Human-Computer Interaction (HCI) for a long time and relatively recently has been started to be study also for videogames. The term has had many different definitions across the time and no general consent has been achieved yet. Among different definitions of UX we could consider:

> From an emotional experience of products point of view, Hekkert describes user experience as "the entire set of affects that is elicited by the interaction between a user and a product, including the degree to which all our senses are gratified (aesthetic experience), the meanings we attach to the product (experience of meaning), and the feelings and emotions that are elicited (emotional experience)" (2006).
>
> From the usability field, Shedroff defines it as "the overall experience, in general or specifics, a user, customer, or audience member has with a product, service, or event. In the Usability field, this experience is usually defined in terms of ease-of-use. However, the experience encompasses more than merely function and flow, but the understanding compiled through all of the senses" (online).
>
> From a computational and technological point of view Hazzenzahl and Tractinsky refer to UX as "a consequence of a user's internal state (predispositions, expectations, needs, motivation, mood, etc.), the characteristics of the designed system (e.g. complexity, purpose, usability, functionality, etc.) and the context (or the environment) within which the interaction occurs (e.g. organizational/social setting, meaningfulness of the activity, voluntariness of use, etc.)" (2006)
>
> Finally, from a web interface design perspective Goto defines UX as "the overall perception and comprehensive interaction an individual has with a company, service or product. A positive user experience is an end-user's successful and streamlined completion of a desired task" (2004).

Note that all the definitions mentioned above include user's emotions and perceptions as the core concept but, however, they vary in other aspects. To Shedroff one important issue, although not the most important, refers to the ease-of-use giving importance to the usability of the product, while Hekkert only takes into account subjective experiences felt by the user. Moreover, Goto mentions the "successful and streamlined completion of a desire task" focusing the UX to the effort done by the user to finish an assigned action. Finally, Hassenzahl & Tractinsky include the term context in their

definition, thus giving importance to social issues that might be concerned with the system.

There are several factors that influence a person's UX with a system and that should be considered when defining the term: "the context – a mix of social, physical, task and technical and information contexts-, the user – the person's states affect the experience-, and the properties of the system" (User Experience White Paper). However, to take into account all the contextual, personal and technical factors when designing UX's can be very slow and sometimes impossible. To achieve the best UX design it is important to employ the most relevant factors to the task to be developed and evaluate them to determine the accuracy of the design.

Finally, it is also important to understand that UX may be subject to the duration of the experience. One can identify even four different types of UX based on the length of these: Momentary, the specific changes during interaction; episodic, the judgment of time of a specific usage episode; cumulative, the whole experience after usage; and anticipated, the previous knowledge the user can have of the system or similar systems.

2.2 Evaluating the User Experience

UX evaluation has been done for years with the goal of identifying the possible system's errors in terms of usability, engagement, and different kinds of emotional states while using a system. There are several different methods for evaluating UX that can be applied in different ways depending on what the developer wants to assess. Here we will identify two different variables: Subjective and Objective methods. These two types of methods can be used either alone or mixing several of them.

2.2.1 Subjective evaluation methods

Subjective assessment methods are mainly used to determine how users behave when interacting with a system. The importance of these methods lies in the subjective feedback, user's perception, given to researchers and evaluators. Subjective assessment data also named self-reported data, can be collected at different times of the evaluation session: at the end of each task (post-task ratings) and at the end of the entire session (post-study ratings) (Tullis & Albert, 2008). Also, there are techniques that allow data collection during the tasks. In the following paragraphs we will describe some subjective assessment methods typically used by researchers.

a) Observation

Observation techniques are one of the most used evaluation methods. Observation gives insights of how the person is dealing with the task by analyzing facial expressions and body language (Mirza-Babaei et al, 2011). Observation methods have to be well

structured by having the goals clear so the researchers can get valuable information from the process.

Moreover, it is highly recommended to record the sessions in video since it captures both visual and audio information (Rogers et al, 2011). However, observation methods tend to be very time-consuming because there is so much to take into account. Also, the quantification of the data is not precise most of the times since it depends on the evaluator to determine the states of the user.

b) Thinking aloud

Thinking aloud is a technique developed by Erikson and Simon (Erikson & Simon, 1985) that consists in asking users to say out loud everything what they think and try to do. This way the process is externalized and the researchers can get data from it. However, thinking aloud techniques tend to interfere with the task the user is developing since it is unnatural (Mirza-Babaei et al, 2011).

c) Questionnaires and Interviews

Questionnaires are very useful for UX evaluation tasks. They are often used since they allow fast and convenient analysis of data. There are several ways to do questionnaires with open-ended or closed questions. To make the analysis process faster it is highly recommended to employ closed questions using Likert scales so the statistics and reports can come easily (Tullis & Albert, 2008).

Although it is true that subjective assessment methods give important insights about user's perceptions it is also important to take into account the different biases that these measures can have (Rogers et al, 2011). The evaluator has to take into account that the perception commented by the user after the task is omitting details of the experience itself and can be subjected to a lack of memory of the user or it can be just the reflect of the "finished" experience. Moreover, when asking to participants there always exists the social desirability bias where participants tend to give answers to the questions thinking of what the researcher is expecting them to say (Tullis & Albert, 2008).

Table 1 describes the pros and cons of the subjective assessment methods mentioned above.

Subjective Measures			
Measure	**Description**	**Pros**	**Cons**
Observation	One of the most used techniques	Allows interpreting facial and body language	Time consuming
		Visual and auditive information	Depends on the subjective insterpretation of the researcher
Thinking Aloud	Users externalize they are doing while they do it	Real-time	Unnatural for the user
Questionnaires and Interviews	Open-ended or Closed-ended. Are highly used in HCI UX evaluation	Give insights of the subjective experience	Reflect the finished experience
		Easy to analyse	Lack of memory of the participant
			Social desirability

Table 1 Summary for the subjective assessment methods

2.2.2 Objective evaluation methods

On the other hand, objective methods refer to those measures that are not dependent on human perception but that are captured and interpreted by a system. These types of methods have the advantage that they are captured in real-time and *they are not contaminated by participant answering style, social desirability, interpretation of questionnaire item wording or limits of participant memory, nor observer bias* (Kivikangas et al., 2010). These methods also vary in their nature since they can capture internal or external states of the person.

a) Non-Physiological objective methods

The most commonly external methods used are log files and pressure sensors. These types of measure allow researchers to identify different cognitive and attentional processes.

Furthermore, as the objective of our research is focused on psychophysiological measures, we will describe different experiments and results obtained by others in this field. However, physiological data is still hard to interpret since it is not a one-to-one relationship, which makes it complicated to create a correlation between physiological measures and psychological states (Nacke, 2011).

As previously said, there are several attempts to use psychophysiological measures for evaluating UX. Here we will describe some of them.

b) Types of Physiological measures.

Eye Tracking

The Eye tracking technique is used to determine where the user is looking at. Through eye tracking measures researchers can recognize fast movements, fixations or pupil dilatation experienced by the user during a determined task (Nacke et al, 2010). Fixations, for example, "represent the instances in which information acquisition is and processing is able to occur" and pupil dilatation is "typically used as a measure of to gauge an individual's interest or arousal in the context they are viewing" (Granka et al, 2004).

Figure 1 Heat map from eye tracking

Heart Rate (HR)

Probably the best known of physiological measures, HR has been largely used as a measure of arousal in many different experiments of psychophysiological measures for evaluating UX. HR measures have been proved to be correlated with different states of affect: high HR indicates feelings of frustration and tension while low HR indicates feelings of competence and immersion (Drachen et al. 2009).

Different kinds of approaches have been done using HR measures. In the field of affective videogames a Wii-based prototype described by Pope & Stevens (2011) is used to translate HR rhythm into a disruption of control of a surgical instrument. Moreover, Janssen et al. (2011) reported findings that demonstrate that listening to other's HR is perceived as an intimate cue due to the intrinsic relation it has with emotions. More examples of HR measurement applications can be found in Mandryk & Atkins, 2007; Martínez et al., 2011; Gilleade & Lee, 2011; Pfurtscheller et al. 2010.

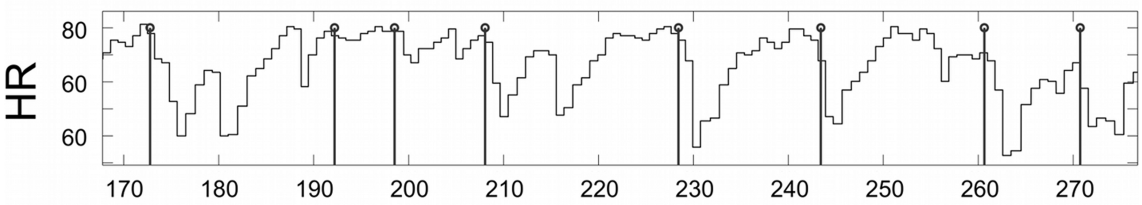

Figure 2 Heart Rate measurement (Pfurtscheller et al, 2010)

Skin Conductance

Skin Conductance - also known as Galvanic Skin Response (GSR), Electrodermal Activity (EDA) or Skin Conductance Response (SCR)- is a measure of the changes of the eccrine sweat glands when they are activated by the central nervous system (Drachen et al, 2009). This type of measure has been largely and successfully used to determine states of arousal in mood induction and videogames tasks.

For example, Mandryk & Atkins (2007) used an Arousal-Valence space created by P.J. Lang in 1995 (Lang, 1995) to classify emotions in a 2D space. They used GSR and Heart Rate to generate an arousal value to fit it in the AV space when participants played a videogame. Finally, they found a linear correlation between arousal and GSR. Furthermore, Drachen et al. (2009) found strong correlations between the Negative Affect dimension of an iGEQ survey, although their results were not linear as in Mandryk & Atkins' experiment (for more examples of Skin Conductance in UX see Mirza-Babaei et al. 2011; Martínez et al. 2011; Kuikkaniemi et al. 2010; Janssen et al. 2011).

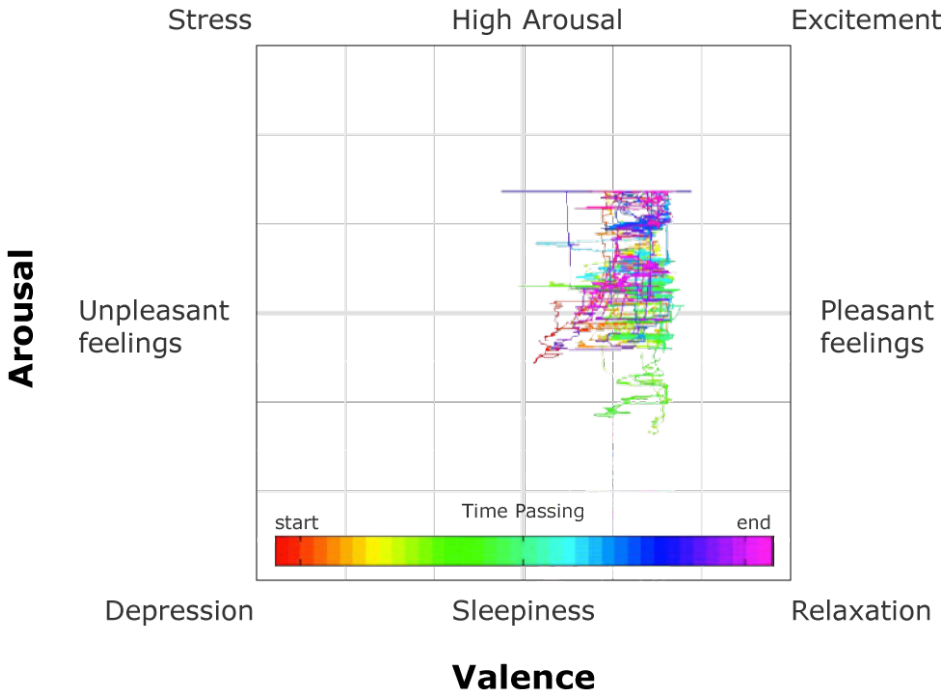

Figure 3 Arousal-Valence Grid used by Mandryk & Atkins, 2007

Facial Electromyography (EMG)

Facial EMG measures the activity of facial muscles. It is commonly used to evaluate positive and negative valence (Kivikangas et al., 2010). It is similar than video coding but it detects even minuscule responses and can give a temporal precision of milliseconds. Although this measure is extremely sensitive to noise it has been largely used in UX research. Mandryk & Atkins (2007) used EMG to detect if the participant was smiling or frowning. With this data they could map the level of valence the participant was experiencing while playing a videogame.

Figure 4 Kuikkaniemi et al Game interface using EMG (2010)

Electroencephalography (EEG)

EEG is the recording, from the scalp, of bioelectrical impulses generated by brain activity; a "sum of microdipoles representing pyramidal cells oriented perpendicular to the surface of the head" (Kropotov, 2009). This type of measure has been largely used in medical applications with the aim to discover brain diseases. For years EEG measures have been used for medical and military applications and recently these have also found a place in Human-Computer Interaction (HCI) and videogames industries through Brain-Computer Interfaces (BCI).

Different types of applications have been developed for different uses since providing handicapped persons with communication and movement skills to training military personnel. BCI applications have been also researched to provide different types of interaction with systems. Tangerman et al. (2009) developed a BCI-based pinball where users had to control a pinball machine through brain activity. Their results showed significant data when comparing users' results to random or non-controlled results.

Some researchers have also focused on using EEG measures to evaluate affective states with some degree of success through mood induction experiments (Wingrave et al., 2011) through the manipulation of EEG rhythms such as Alpha, Beta and Theta. Furthermore, BCI systems can be applied to UX research by adapting the interface or videogame based on the goal of keeping the user in a state of "flow" (Niholt et al., 2009).

However, although the level of success reported by different authors varies depending on the application, it is clear that the optimal signal processing and analysis for BCI systems is still to be achieved. EEG measures are too sensitive to noise and misunderstandings due to psychophysiological inference since it is not usually a one-to-one relationship between physiological measure and psychological state (Nacke, 2011).

Figure 5 Beta rhythms from EEG (Pfurtscheller et al, 2010)

Table 2 describes the pros and cons of the objective assessment methods mentioned above.

Objective Methods			
Measure	**Description**	**Pros**	**Cons**
Non-Physiological Measures			
Log Files	Record the information of the navigation of the user when using the system	Record information in Real-time	Difficult to analyse
		Precise	Long Files
Physiological Measures			
Heart Rate (HR)	The measure of the beats of the heart	Correlated with arousal	Lack of expressiveness
Skin Conductance (SC)	The measure of the changes of the eccrine sweat glands	Correlated with arousal and valence	Noise
		Negative Affect	
Facial Electromiography (EMG)	The measure of the activity of the facial muscles	Precision of milliseconds	Highly noisy
		Correlated with valence	
Electroencephalography (EEG)	The measure of bioelectrical impulses generated by brain activity	New way of interaction with systems	Highly noisy
		Games can react to user's brain activity	Hard to interpret since there is no one-to-one correlation between psychological and physiological states

Table 2 Summary for the objective assessment methods

2.3 Current state of physiological measures applied to UX evaluation

As reported in the previous review of UX evaluation methods - both subjective and objective - we can see how there is an attempt to develop a framework that allows researchers to achieve standard measures of UX. Subjective assessment methods have been widely used for HCI research and have been, somehow, standardized although

there is still no general agreement about how to use them for each researcher has to define the best way to apply it to their experiments. However, for physiological measures there is still no agreement about how to use them either about how to interpret them. These techniques are being widely used in game research and there are improvements in their interpretation, but we still need to find the specific correlations between physiological measures and human perception that allows us to use them separately and collect meaningful data to support our investigation.

a) Biometric Storyboards

Biometric Storyboards is a UX evaluation framework developed by Pejman Mirza-Babaei and Graham McAllister (2011). This framework aims at integrate both psychophysiological and subjective assessment methods. In their experiment, they implemented GSR, video tapping and streaming of the game. After the experiment session they developed the storyboard showed in Figure 1 (the images of the streaming of the game were removed because the game was unreleased). With green and red points they marked the arousal state of the player and related it to a moment of the game. Afterwards, they could analyze the most important and engagement stages of the game.

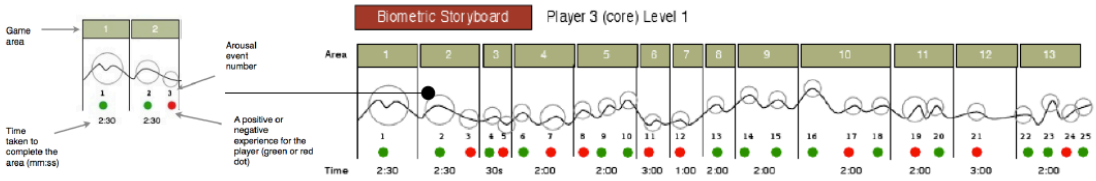

Figure 6 Example of a biometric storyboard

This approach provides a new way for understanding UX evaluation since it relates in a storyboard psychophysiological measures with observational methods. It allows researchers to understand the failures and the successes of the product and gives an insight of the user's perception unlinked to the social desirability or other problems of subjective self-reported measures.

3. Multimodal framework

Based on the State of the Art we have developed a multimodal framework that combines physiological measures (EEG) and questionnaires in order to assess UX. With this framework we want to improve traditional, self-reported UX evaluation methods by adding EEG measures that can remove or at least reduce biases that are implicit in subjective, self-reported measures.

In the first place it was important to define the questionnaire that was going to be used as the self-reported measure. We decided to create a new 5-point Likert scale questionnaire combining two different questionnaires in order to assess three variables that could be effectively correlated with EEG measures. Those variables are: Challenge, Enjoyment and Excitement. In the variable of Challenge we found it appropriate to use the Game Experience Questionnaire (IJsselsteijn et al, unpublished). This questionnaire has been used in different studies towards evaluating UX (see Drachen et al, 2010). On the other hand, the variables of Enjoyment and Excitement were extracted from a questionnaire developed by Giakoumis (2009). This questionnaire was used in similar conditions as the topic of this framework. See Appendix 1 to access to the full questionnaire.

3.1 Variables and Components

Each variable has from five to six components that describe different aspects of the variable to be analyzed. The "Challenge" variable has six components in total that represent three categories in terms of "Effort", "Learning" process and "Time Pressure" perception. These components allow us to identify the level of perceived challenge during each task.

The variable of "Enjoyment" has five components, two of them shared with the variable of "Excitement": 'I enjoyed' and 'I felt excited'. These two components are important to analyze both variables since both of them can assess levels of valence and arousal. The components of this variable analyze on the one hand the levels of "Pleasure" experienced during the task and, on the other hand, the perception of the dominion of "Knowledge" that allow them to complete the tasks.

Furthermore, the variable of "Excitement" has five components that analyze both, negative and positive aspects of arousal. Firstly, we have questions that ask for positive aspects such as "Excitement", "Enjoyment" and the "Impressiveness" of the experience. And secondly we ask participants about their levels of "Boredom" and "Irritability" during the tasks.

Further, we decided to work with "Alpha rhythms" – it means with frequencies in the range from 8 to 13Hz – present in the EEG signal. Activity within the alpha band is associated with states of arousal and relaxation (Kropotov, 2009). See Figure 7 for a graphical representation of the variables and components included into the framework.

Finally, in order to this framework to work, a synchronization process and a correct preparation of tasks must be done. Synchronization between tasks and EEG recordings is a fundamental issue to be able to recognize, during the task, the emotional states of the participant across time. Therefore tasks must have a beginning and an end that allow researchers to determine the exact point during the task in which a peak in the Alpha rhythm was identified. Moreover, to test the framework we prepared four tasks that went in an ascendant degree of difficulty in order to take the participants through different states of challenge, enjoyment and excitement.

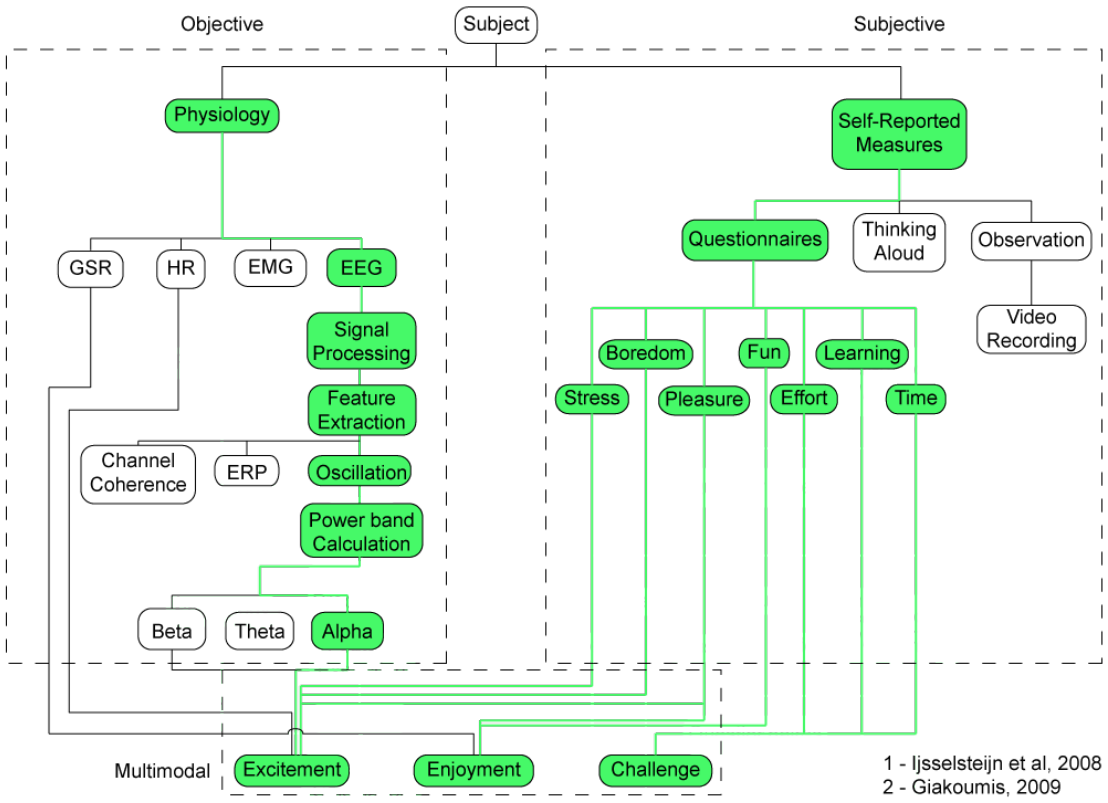

Figure 7 Diagram of the Multimodal Framework (in green the parameters being used)

4. Methods

For this thesis we embrace tabletop interfaces as a case study for the proposed multimodal framework. In our experiments we used reactable (Jordà et al, 2007) as the interface with which participants interacted and for which the tasks were prepared. In this regard we propose a number of experiments for testing the feasibility of EEG - in particular Alpha rhythms- as an assessment method for UX evaluation, an experimental design for tabletop systems and a specific protocol to use the multimodal framework.

4.1 Technical aspects

For our experiments we used principally two hardware systems: Reactable and Emotiv EPOC. Here we will describe some technical aspects about these systems.

a) Reactable

The reactable is a multi-user tabletop tangible user interface for music creation. Reactable has different types of objects that allow the user to generate, filter, control and mix sounds. Moreover, there are objects that allow the user to control global parameters such as tempo and volume. The reactable also allows the users to interact with the tactile surface where the objects are placed.

Figure 8 Reactable (Jordà et al, 2007)

b) Emotiv EPOCH

The Emotiv EPOC is a high resolution, neuro-signal acquisition and processing wireless neuroheadset. It allows monitoring engagement, boredom, excitement, frustration and meditation level in real time. It has 14 channels with a sampling rate of 128Hz and electrodes positioned at AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8 and AF4 (the international 10-20 system).

Figure 9 Emotiv EPOCH

4.2 Pilot Experiment

The pilot experiment was executed using the Reactable. This experiment was thought to determine whether the mere fact of using a BCI headset biases the response of subjects during UX evaluation in terms of social desirability and engagement.

a) Sample

For the pilot experiment, 12 participants, 6 males and 6 females, with ages between 23 and 36 years old ($M = 27.50, SD = 3.966$) were recruited via email.

b) Groups

Participants worked in couples to carry out the experiment. Each couple performed two tasks where they used the Emotiv EPOC for the first task and they did not use it for the second task. However, in order to obtain feasible results the order of the tasks was randomized. It is important to remark that the second task was slightly harder than the first task.

c) Measures

Before the beginning of the tasks participants filled a demographic questionnaire where they were asked for their age, sex, nationality, music knowledge and previous

experience with reactable. Also, as shown in Figure 7 we used two post-task questionnaires, both of them were exactly the same since the aim of this experiment, as previously mentioned, was to analyze significant differences between subjective self-reported measures under two conditions: wearing and not wearing a BCI. The questionnaires were focused on three different variables: Challenge, Enjoyment and Excitement. All the questions were randomized through the questionnaire in order to obtain more accurate results. Also, we provided a SAM (Self-Assessment Manikin) – scale (Lang et al, 1995) questionnaire before the experiment started and another when the experiment finished.

d) Task

Before starting, participants fulfilled a SAM scale questionnaire followed by a demographic questionnaire. Further they received an explanation about the characteristics of Reactable followed by two minutes when they explored with all the objects available for the experiment. Participants were asked to perform, collaboratively, two tasks of replication of sounds. Figure 10 shows the experimental protocol as the timeline participants followed.

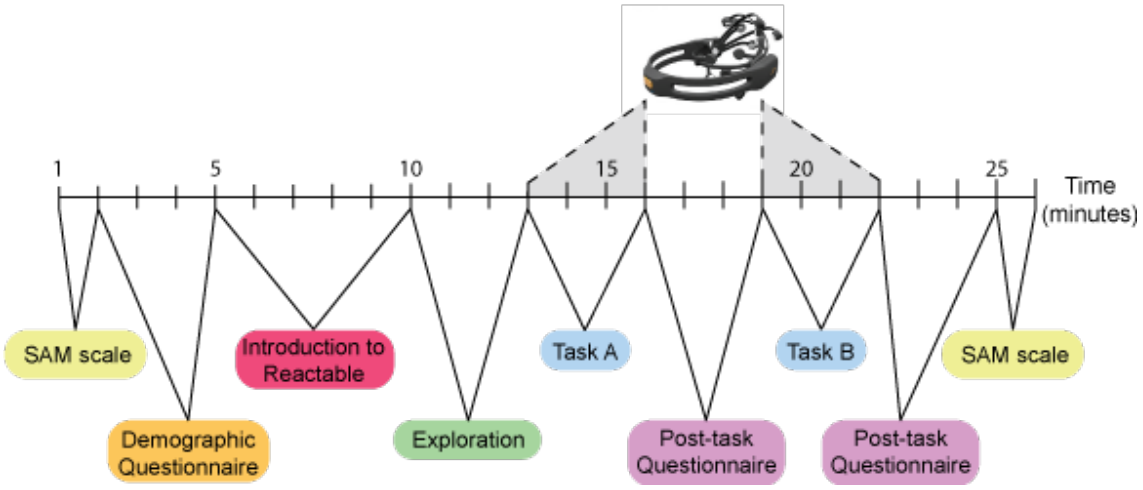

Figure 10 Experimental protocol for the Pilot Experiment

4.3 Experiment

This experiment was designed to test possible correlations between physiological signals (EEG) and subjective measures, questionnaires.

a) Sample

Here, 21 subjects, 6 women, were recruited via email. The mean age of the participants was 30.38 years, std. dev. = 4.85.

b) Groups

All the participants followed the same experimental protocol so groups and conditions did not exist in this experiment.

c) Measures

Before the beginning of the tasks participants filled a demographic questionnaire where they were asked for their age, sex, nationality, music knowledge and previous experience with reactable. During the experiment we measured EEG activity using the Emotiv EPOC hardware, which we placed right before the experiment began. For subjective, self-reported measures we used the 5-point Likert scale questionnaire previously described – with the same variables of challenge, enjoyment and excitement – to evaluate each task. All the questions were randomized through the questionnaire in order to obtain more accurate results. Also, we provided a SAM – scale questionnaire before the experiment started and another when the experiment finished.

d) Task

The experiment was designed for the reactable system. As previously mentioned, we prepared four tasks that went in an ascendant degree of difficulty in order to take the participants through different states of challenge, enjoyment and excitement. First, participants followed a simple tutorial where we explained the way reactable works and which were the properties of each of the objects they were going to use.

Further, they carried out four imitation-of-sound tasks. Figure 11 shows the experimental protocol.

Each task required new reactable objects that participants had to combine to generate the sound. Every sound that was generated in a task was used in the following task.

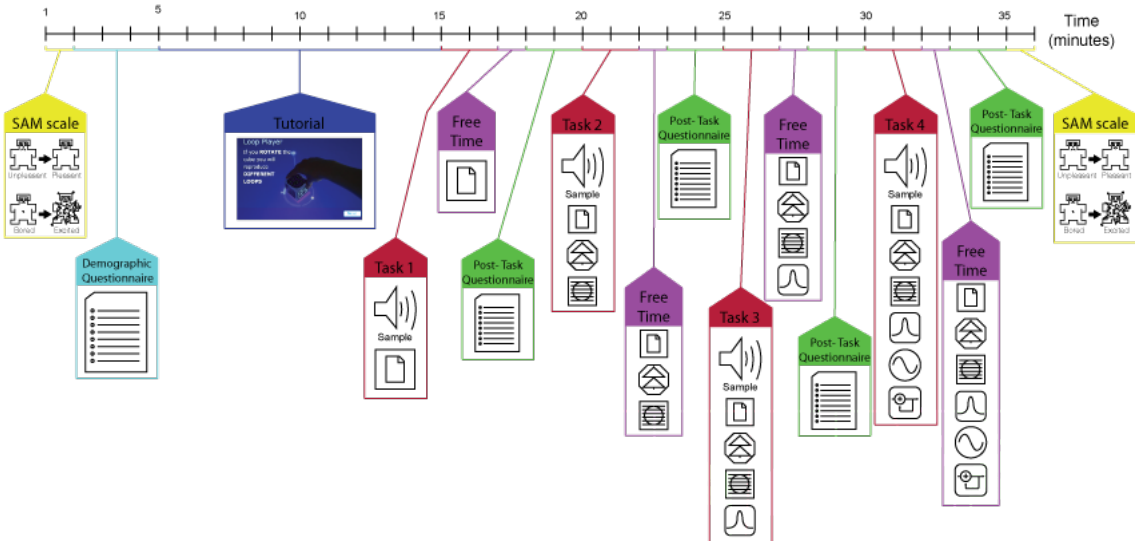

Figure 11 Experimental protocol for the experiment

5. Results

5.1 Pilot Experiment

An Analysis of Variance (ANOVA) was performed to detect significant differences in each variable for both conditions. The analysis showed that the means for all variables were higher for the non-BCI condition when compared to the BCI condition, but not at a significant degree (see Figure 12).

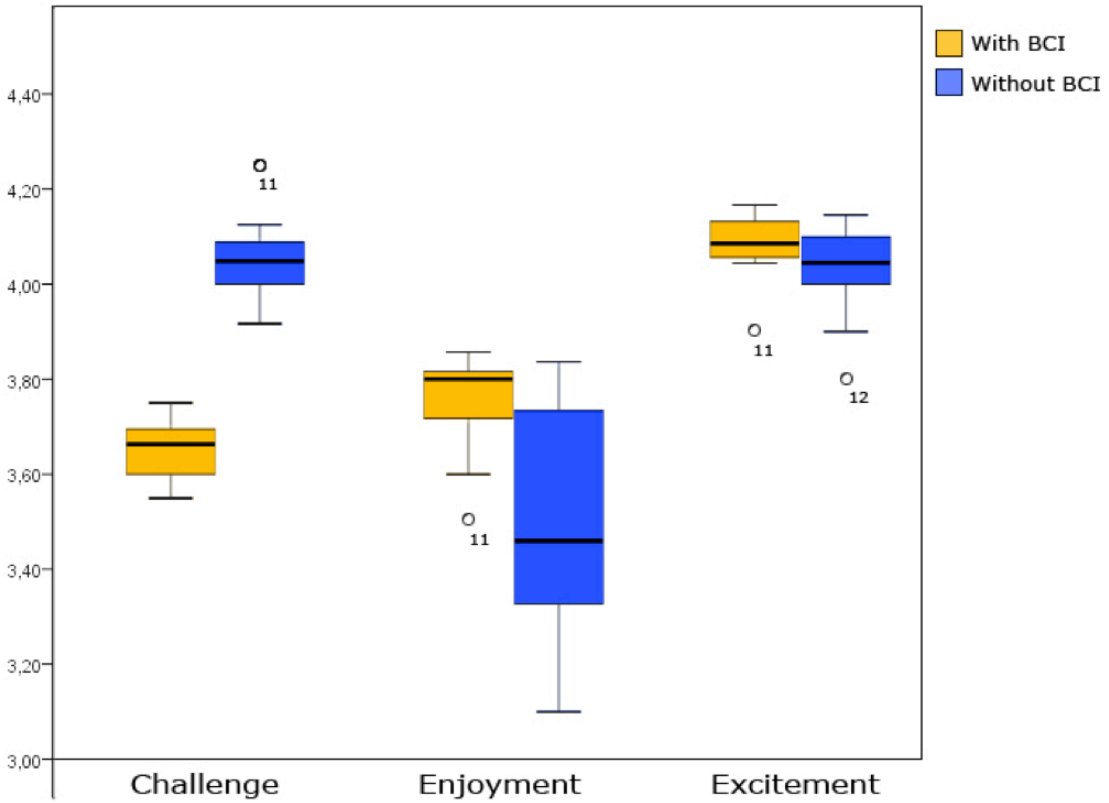

Figure 12 Boxplot for variables across conditions

a) Analysis by component

After determining that there were no significant differences between, we went deeper to study the differences for each component of the variable. No significant differences were found between the two conditions, although for the questions "*I felt that I was learning*" and "*I thought it was hard*" we found a close-to-significant result ($p = 0.059$ and $p = 0.053$ respectively). This result might be related with the fact that Task 2 was slightly harder than Task 1.

b) Correlations

Moreover, we expected positive correlations to exist between conditions in the same variable since the tasks had a difficulty difference. For all the three variables we found positive correlations although none of them was significant.

Further, to test the regularity of the questionnaire different correlation coefficients were calculated. First of all, we tested the correlations between questions in the same variable.

Challenge

For this variable all the correlations tended to be positive, although not all of them were strong. However, two pairs were negative. The first one we will discuss is the following pair: I felt stimulated – I thought it was hard ($r = -0.243$). This negative correlation can also be found when comparing these two components only under the condition where participants were not wearing the BCI headset ($r = -0.578$). Checking the means for these two questions – $M = 4.04, SD = 0.908; M = 3.46, SD = 0.977,$ respectively - we realized that the easier the task was perceived the more stimulated the participant was. This could be explained due to a lack of knowledge about reactable where participants felt more stimulated when developing easier tasks that they could complete.

The second negatively correlated pair was: I felt that I was learning – I felt challenged ($r = -0.166$). If we take a deeper look into the data, we find that it is only negatively correlated for the condition where participants were not wearing a BCI headset. Taking into account the different correlation coefficients between conditions we will assume that since the second task was more difficult than the first, participants felt the task more as a challenge than a learning process.

Enjoyment

In this variable we found that the correlation coefficient for all the pairs of components, except for one, were positive. The negatively correlated pair was: I felt I had enough previous knowledge to complete the task – I enjoyed ($r = -0.155$). According to this result we could assume that the less they knew about reactable the more they enjoyed playing with it.

Excitement

In this variable we found similar results than the other two variables. Only two of the correlation coefficients do not correspond to what was expected. These components are: I found it impressive – I felt bored and I found it impressive – I felt irritable. In these two cases the correlation coefficients expected were negative, however we found strong and positive correlation coefficients – ($r = 0.522$ and $r = 0.344$ respectively). Looking at the data we envisioned that this problem is probably related to the small amount of participants since the differences in the means are highly significant ($p < 0.001$).

5.2 Experiment

As described in section 4, the proposed framework has been designed to increment challenge, enjoyment and excitement across tasks, as well as activity in the EEG Alpha band.

In order to analyze the data obtained through the Emotiv EPOC we extracted Alpha power band (7-13hz), Beta power band (13-30hz) and Theta power band (4-8hz) from the T8, T7, P8, P7, O1 and O2 channels. To be able to process the data and extract the frequencies we were willing to analyze we used MATLAB and the EEGlab plug-in[1]. However, no significant results were found for Beta and Theta power band analysis. In the next section we will describe the results obtained for the Alpha power band analysis.

5.2.1 Questionnaire results

a) Significance tests

First of all, we used a T-test to look for significant differences between the results for Challenge variable depending on the task – Task 1 (*M = 3.36, SD = 0.60*), Task 2 (*M = 3.81, SD = 0.69*), Task 3 (*M = 3.38, SD = 0.63*) and Task 4 (*M = 4.02, SD = 0.67*). As expected when seeing Figure 13, there exist significant differences for the Challenge variable between tasks 1 and 2 ($p < 0.01$), 1 and 3 ($p < 0.01$), and 1 and 4 ($p < 0.001$). This result allows us to assert that tasks were actually perceived as harder across time, and the difference between the first and the last task was the desired. However, no significant differences were found for the variables of Enjoyment and Excitement.

[1] http://sccn.ucsd.edu/eeglab/

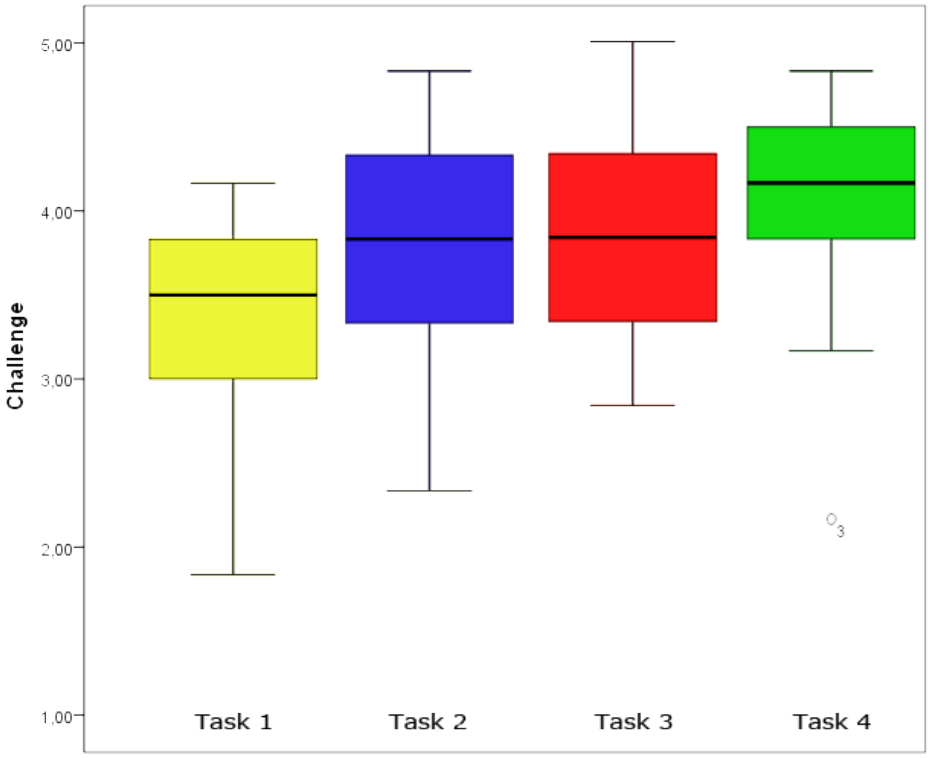

Figure 13 Boxplot for Challenge across tasks

To gather more information we did a one-way ANOVA with a LSD correction in order to test for differences in each component of variables between tasks. In the variable of challenge we found that components that were asking for effort levels – such as 'I thought it was hard' and 'I had to put a lot of effort into it' - showed significant differences across tasks. Although most of the analysis did not show significant differences at all, two questions are interesting to analyze.

I thought it was hard – Variable: Challenge

In this component we found significant results between Tasks (Task 1, $M = 2.81$, $SD = 0.281$; Task 2, $M = 3.81$, $SD = 0.814$; Task 3, $M = 3.86$, $SD = 1.1014$; Task 4, $M = 4.14$, $SD = 0.854$). Further, the significant differences were found in Task 1 and 2 ($p < 0.01$), Task 1 and 3 ($p < 0.001$) and Task 1 and 4 ($p < 0.001$). This result was already previewed since the level of difficulty of the tasks was ascendant.

I had to put a lot of effort into it – Variable: Challenge

In this component there also exist significant differences in the results and they show how the easiest task was the first one (Task 1, $M = 2.86$, $SD = 1.014$; Task 2, $M = 3.76$, $SD = 0.889$; Task 3, $M = 3.9$, $SD = 0.889$; Task 4, M 0 4.14, $SD = 0.854$). We found significant differences between Task 1 and 2 ($p < 0.01$), Task 1 and 3 ($p < 0.001$) and Task 1 and 4 ($p < 0.001$). Although there are no significant differences between the other tasks we can notice that the effort done by the user is bigger every time.

I felt time pressure – Variable: Challenge

The results in this question are also related to the previous two questions (Task 1, *M = 2.95, SD = 1.284;* Task 2, *M = 3.76, SD = 1.221;* Task 3, *M = 3.33, SD = 1.111;* Task 4, *M = 3.71, SD = 1.384*). We obtained significant differences between tasks 1 and 2 ($p < 0.05$) and 1 and 4 ($p < 0.05$). Although the time was the same for all the tasks, the difficulty of the task might influence the perception of time during the experiment, so as the task is perceived more difficult the time is perceived shorter.

I felt irritable – Variable: Excitement

In this question there are also significant differences. We found significant differences between task 1 and 2 ($p < 0.05$) and 1 and 4 ($p < 0.05$). However, these differences are not consistent across time since the means were: Task 1, *M = 1.19, SD = 0.512;* Task 2, *M = 2.14, SD = 1.153;* Task 3, *M = 1.76, SD = 1.136;* Task 4, *M = 2.14, SD = 1.236*. Nonetheless, we can notice that irritability increases when two – or more - objects are introduced to reactable in the task.

b) Correlations

Furthermore, we also tested data aiming to find correlations between the different variables – challenge, enjoyment and excitement. Figures 14 to 16 show how that in fact there exists a tendency to be positively correlated between variables. These results allow us to assert that all the variables are somehow related.

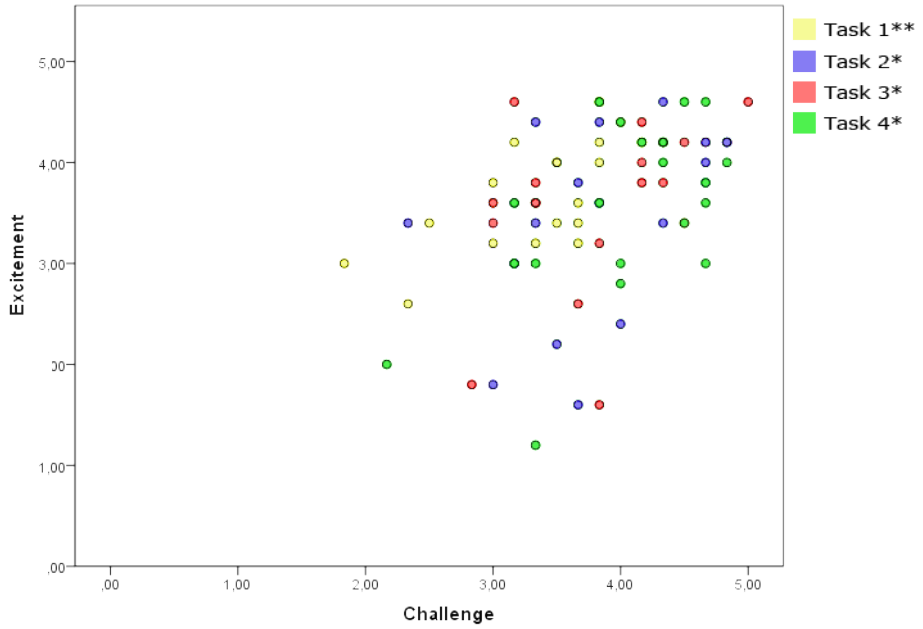

Figure 14 Scatter plot for Challenge and Excitement across tasks

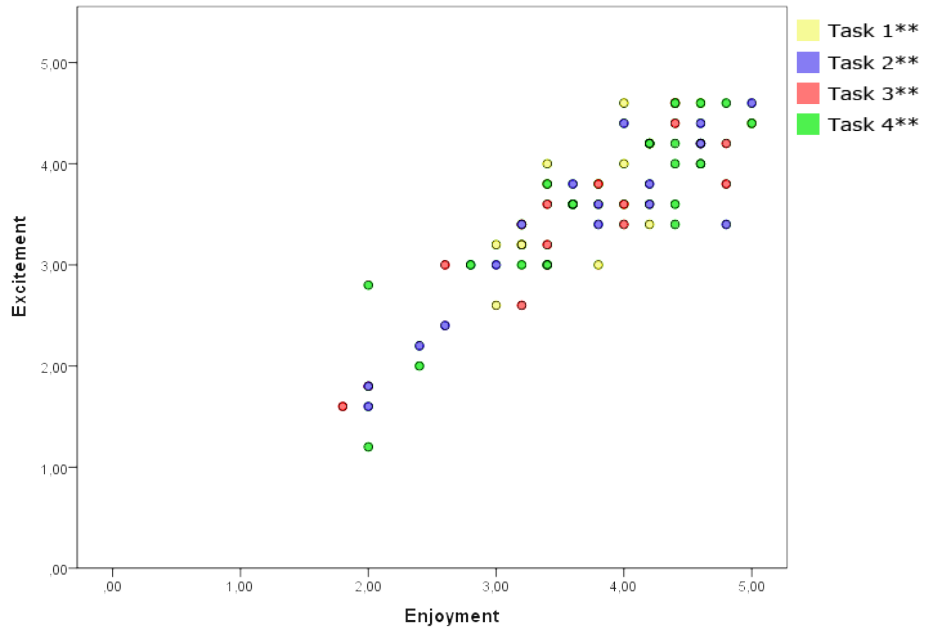

Figure 15 Scatter plot for Excitement and Enjoyment across tasks

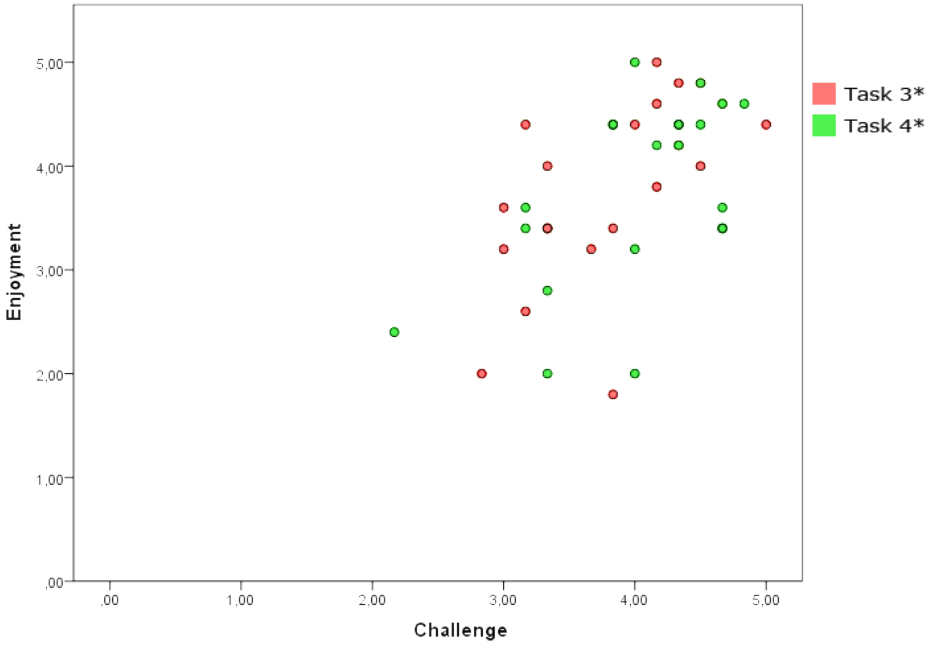

Figure 16 Scatter plot for Enjoyment and Challenge across tasks

In order to understand the lack of correlations between enjoyment and challenge in the first two tasks more analysis were done. First of all we excluded the component *I felt I had enough previous knowledge to complete the task* from the variable of enjoyment and did the correlations again. As shown in Table 9 there exist correlations between enjoyment and challenge for three of the four the tasks.

		Enjoyment 1	Enjoyment 2	Enjoyment 3	Enjoyment 4
Challenge 1	Pearson Correlation	0,502*	0,255	0,335	0,048
	Sig. (2-tailed)	0,020	0,264	0,138	0,837
Challenge 2	Pearson Correlation	0,132	0,340	0,238	0,184

	Sig. (2-tailed)	0,567	0,132	0,299	0,424
Challenge 3	Pearson Correlation	0,680**	0,427	0,557**	0,381
	Sig. (2-tailed)	0,001	0,054	0,009	0,089
Challenge 4	Pearson Correlation	0,502*	0,575**	0,579**	0,447*
	Sig. (2-tailed)	0,020	0,006	0,006	0,042

Table 3 Correlations between Challenge and Enjoyment for the four tasks

Further, we correlated the components of challenge with the new values of Enjoyment. We noticed that there are strong correlations for all the tasks between Learning and Enjoyment – as shown in Figure 17. However, no significant correlations were found for the components of Challenge: Effort and Time.

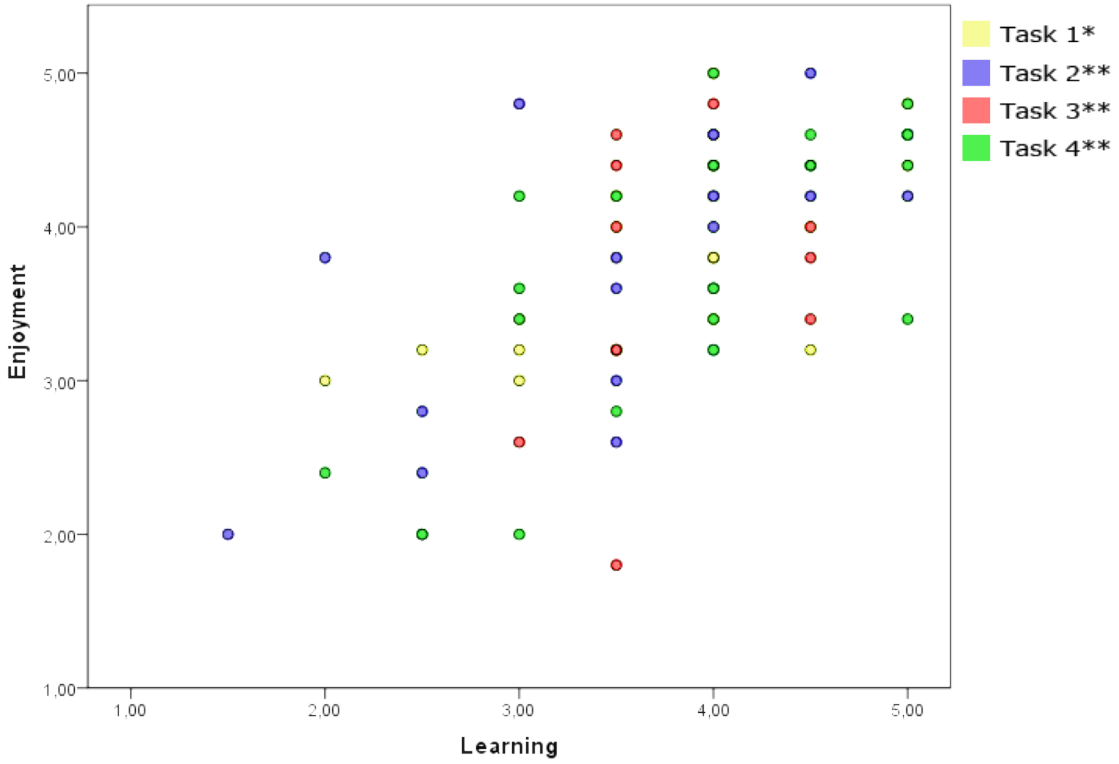

Figure 17 Scatter plot for Enjoyment and the component Learning

5.2.2 EEG results

In a general exploration of data we realized that four subjects presented extreme values and were categorized as outliers and were removed from the analysis. Then we proceeded to statistically analyze the data.

a) Comparison of Means

No significant neither close-to-significant differences were found when analyzing Alpha band activity across tasks (Task 1, *M = 5.42, SD = 2.26;* Task 2, *M = 5.51, SD = 1.57;* Task 3, *M = 5.69, SD = 1.71;* Task 4, *M = 5.69, SD = 2.23*).

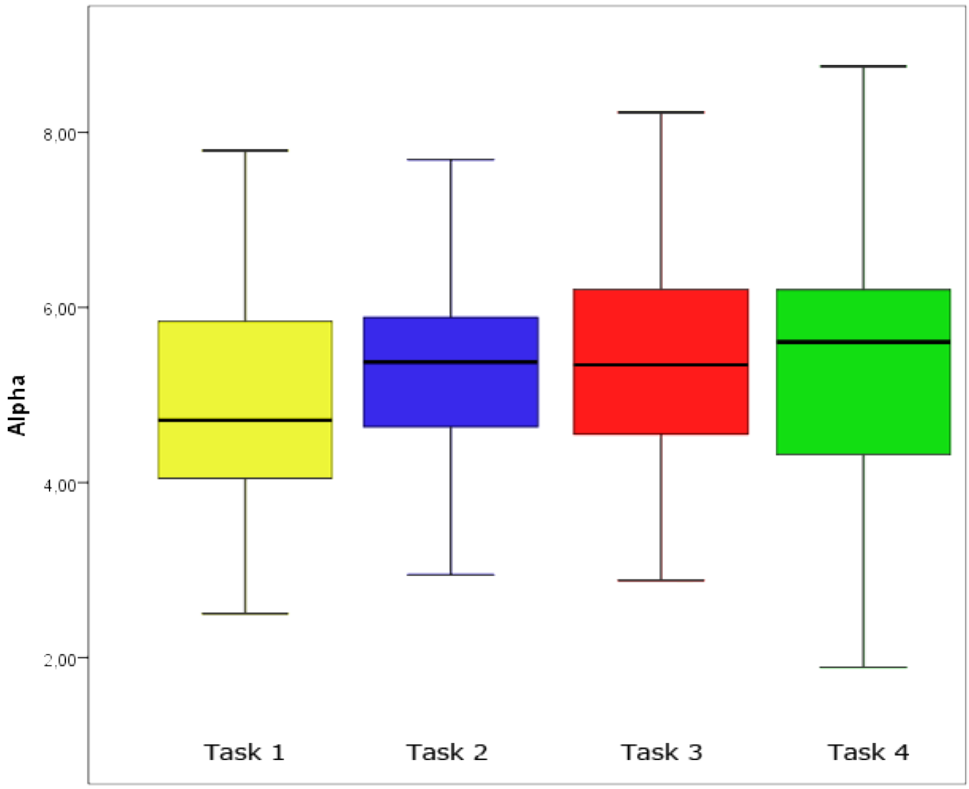

Figure 18 Boxplot for the means of Alpha Power Band across tasks

b) Correlations between questionnaires and EEG

Further, we analyzed the data in the interest of finding correlations between the data obtained with the questionnaire and the alpha power band. The first step towards this process was analyzing the correlations between alpha and the general variables of the questionnaire – Challenge, Enjoyment and Excitement. Table 12 shows the results for the correlations with the three variables of the questionnaire.

	Correlation Coefficient with alpha power band			
	Task 1	**Task 2**	**Task 3**	**Task 4**
Challenge	0,078	0,000	-0,047	0,307
Enjoyment	0,056	0,339	0,300	0,419
Excitement	0,360	0,452*	0,573*	0,616**

Table 4 Correlation coefficient of Alpha power band with the categories in the questionnaire (* < 0.05; ** < 0.01)

We can notice that there only exist significant correlations with the variable of Excitement.

Furthermore, we went deeper into the data to test for correlations of alpha with each of the components of the variables. Below we describe the most important results for each variable.

Excitement

Boredom: Correlations in this component were significant and negative for tasks 3 and 4 (r = *-0,554; r = -0,494* respectively). The first two tasks also show a negative trend although not significant.

	Task 1	Task 2	Task 3	Task 4
Correlation	-0,17	-0,385	-0,554*	-0,494*
Sig.	0,514	0,127	0,21*	0,44*

Table 5 Correlation coefficients between the component of Boredom and Alpha Power Band

Stress: in this component there exist negative and significant correlations for tasks 2, 3 and 4 (*r = -0,500; r = -0,494; r = -0,578* respectively) that are coherent with the correlations with the previous component.

	Task 1	Task 2	Task 3	Task 4
Correlation	-0,286	-0,500*	-0,494*	-0,578*
Sig.	0,266	0,041	0,044	0,015

Table 6 Correlation coefficients between the component of Stress and Alpha Power Band

6. Discussion and Conclusion

6.1 Experiment

With this experiment we found four important effects related with both subjective measures and alpha power band. In this section we will interpret how these effects reinforce traditional UX evaluation.

a) Effect 1 – Accuracy when measuring Effort and Stress

According to the results gathered from the subjective, self-reported measures we found that the framework is accurate when measuring effort and stress in tasks with variable difficulty. As we previously discussed in the Results section the measures from this assessment tool reflected the increasing difficulty with which the tasks were designed.

b) Effect 2 – Enjoyment and Excitement evolve together with Challenge

Another effect found with respect to the subjective, self-reported measures, is the evolution of Enjoyment and Excitement with Challenge. As shown in the Results section, the more challenged the person felt the more they enjoyed and felt excited. This effect is important when developing more experiments since we can control these variables by manipulating the levels of Challenge.

In order to test different relations that might exist between physiological signals and emotional states, this information can turn out to be very useful. When developing new experiments one must take into account this information and be aware of the consequences that might bring a change in one of this variables over the others.

c) Effect 3 – Enjoyment increases when a learning effect occurs

From the subjective, self-reported measures we found that Enjoyment evolves together with Challenge when the analysis is focused on the component of Learning. We can argue that a learning effect is reflected in the levels of enjoyment when developing a task.

This type of interaction between these two variables is an important fact that needs to be taken into account when designing future experiments. It would be important to analyze the levels of perceived challenge that can be acquired by generating a learning effect in a designed task.

d) Effect 4 – Alpha Power Band evolves together with the components of Boredom and Stress

Finally, the most important effect we found is the evolution of alpha power band with the components Boredom and Stress. This effect is important since now we know that alpha power band is a good representation of these states even under eyes opened and moving conditions. Taking into account these results we can use alpha power band as a measure of these states when interacting with different types of interfaces. Therefore, it can be used as a measure when evaluating user experience.

This effect also implies an improvement of the traditional subjective, self-reported measures because it turns Excitement into a multimodal variable. As a multimodal variable we can have a better understanding of the feelings of the user while interacting since it allows removing biases from both subjective and objective measures. Furthermore, as a multimodal variable, we can design different types of tasks that imply multiple activities to be evaluated with a higher level of accuracy.

7. Problems and Future Work

In this Thesis we tested the multimodal framework by analyzing alpha power band and its relation with subjective, self-reported measures when interacting with Reactable. However, there is still more work that needs to be done in order to improve the framework.

First of all we found problems when we were trying to define the subjective, self-reported method, since there is no general agreement in which questionnaire to use for each framework. We selected the above-mentioned questionnaire but still more experiments must be performed using different interfaces in order to test the reliability of the framework. Until now, we have been using Reactable but it is important to use interfaces that require different amounts of cognitive load and that generate different states of flow. Also, the complexity and nature of the tasks must vary in order to include tasks that are not influenced by previous knowledge.

Moreover, the integration of subjective and objective was not trivial either. We propose that, in order to obtain better results from the physiological data, we must include different physiological measures that allow us to process the signals and remove artifacts caused by the experimental conditions. Also, these new measures must allow us to assess the different variables – Challenge, Enjoyment and Excitement – in a multimodal way.

Finally, we also found problems with the signal processing. The amount of artifacts generated by the experimental conditions influences the final results we obtained. In order to deal with this problem, we want to analyze the signals obtained across time in order to detect critical points during the interaction with the system instead of analyzing one value that reflects the complete experience. Also, the generation of a continuous signal will allow us to identify different artifacts that might be biasing the final result.

REFERENCES

Antley, A., Slater, M. *"The effect on lower spine muscle activation of walking on a narrow beam in virtual reality*". *IEE Transactions on Computer Graphics and Visualization (2011): 255-259.*

Drachen, A., Nacke, L., Yannakakis, G., Pedersen, A.L. *"Psychophysiological correlations with game experience dimensions". CHI (2009).*

Drachen, A., Nacke, L., Yannakakis, G., Pedersen, A.L. *"Correlation Between Heart Rate, Electrodermal Activity and Player Experience in First-Person Shooter Games". Sandbox '10 Proceedings of the 5th ACM SIGGRAPH Symposium on Videogames (2010): 49-54*

Giakoumis, D., Vogiannou, A., Kosunen, I., Devlaminck, D., Ahn, M., Burns, A.M., Khademi, F., Moustakas, K., Tzovaras, D. *"Multimodal Monitoring of the Behavioral and Physiological State of the User in Interactive VR Games". eNTERFACE (2009)*

Gilleade, K., Dix, A. "*Using frustration in the design of adaptive videogames". ACE (2004): 228-232.*

Guilleade, K., Lee, K. "*Issues inherent in controlling the interpretation of the physiological cloud". CHI (2011).*

Hassenzahl, M., Tractinsky, N. *"User experience – a research agenda". Behavior & Information Technology. Volume 5, Issue 2 (2006).*

Hekkert, P. *"Design aesthetics: Principles of pleasure in product design". Psychology Science, 48(2), 157-172 (2006).*

Jansen, J., Westerink, J., Ijsselsteijn, W., van de Zwaag, M. *"The role of physiological computing in counteracting loneliness*". *CHI (2011)*

Goto, K. "*Brand value and the user experience". The web professional's online magazine of choice (2004)*

Granka, L.A., Joachims, T., Gay, G. *"Eye Tracking Analysis of User Behavior in WWW Search". SIGIR '04 (2004)*

Kropotov, J.D. *"Quantitative EEG, Event Related Potentials and Neurotherapy", Elsevier, Amsterdam (2009)*

IJsselsteijn, W. A., de Kort, Y. A. W. & Poels, K. *"The Game Experience Questionnaire: Development of a self-report measure to assess the psychological impact of digital games". In preparation.*

Jordá, S., Geiger, G., Alonso, M., Kaltenbrunner, M. *"The reacTable: Exploring the Synergy between Live Music Performance and Tabletop Tangible Interfaces". TEI '07*

Proceedings of the 1st international conference on Tangible and embedded interaction (2007): 139-146

Kivikangas, J. M., Ekman, I., Chanel, G., Järvelä, S., Cowley, B., Salminen, M., Henttonen, P., Ravaja, N. "*Review on psychophysiological methods in game research*". *Proc. Of 1st Nordic DiGRA (2010)*

Kuikkaniemi, K., Laitinen, T., Turpeinen, M., Saari, T., Kosunen, I., Ravaja, N. *"The influence of implicit and explicit biofeedback in First-Person Shooter games*". *CHI (2010): 859-868.*

Lang, P. J. *"The emotion probe". American Psychologist 50 (5) (1995): 372-385.*

Mandryck, R., Atkins, M. *A "fuzzy physiological approach for continuously modeling emotion during interaction with play techniques". International Journal of Human-Computer Studies 65, 4 (2007): 329-347.*

Martinez, H., Garbarino, M., Yannakakis, G. *"Generic physiological features as predictors of player experience". Affective computing and intelligent interaction (2011): 267-276.*

Menon, V., Jayaraman, B., Govindaraju, V. *"Biometrics driven smart environments: Abstract framework evaluation*". *UIC 2008, LNCS 5061 (2008): 75-89.*

Mirza-Babaei, P., McAllister, G. *"Biometric Storyboards: Visualising meaningful gameplay events". CHI (2011): 2315-2320.*

Mirza-Babaei, P., Long, S., Foley, E., McAllister, G. *"Understanding the contribution of Biometrics to games user research*". *Authors & Digital Games Research Association DiGRA. Proceedings of DiGRA 2011 Conference: Think Design Play (2011)*

Nacke, L., Drachen, A., Göbel, S. "*Methods for evaluating gameplay experience in a serious gaming context*". *Journal of Computer Science in Sport 9, 2 (2010)*

Nacke, L. "*Directions in physiological game evaluation and interaction*". *CHI BBI (2011)*

Nijholt, A., Bos, D. P. O., Reuderink, B., *Turning shortcomings into challenges: brain-computer interfaces for games. Entertainment Computing (2009): 85-94*

Pfurtscheller, G., Allison, B., Brunner, C., Bauemfeind, G., Solis-Escalante, T., Scherer, R., Zander, T., Mueller-Putz, G., Neuper, C., Birbaumer, N. "*The Hybrid BCI*". *Frontiers in Neuroscience 4, 42 (2010)*

Pope, A., Stephens, C. *"'Movemental': Integrating movement and the mental game". CHI (2011)*

Rogers, I., Sharp, H., Preece, J. "*Interaction design: beyond Human-Computer Interaction*". *Chichester: Wiley (2002)*

Shedroff, N. *An evolutionary glossary of experience design*, online glossary at http://www.nathan.com/ed/glossary/ (23.5.2006).

Tahn, C., Subramanian, S. *"Online single trial ERN detection as an interaction aid in HCI applications*". *CHI (2011)*

Tangerman, M., Krauledat, M., Grzeska, K., Sagebaum, M., Vidaurre, C., Blankerts, B., Müller, K. R. *"Playing pinball with non-invasive BCI"*. *Advances in Neural Information Processing Systems, Vol. 21, (2009): 1641-1648*

Wingrave, C., Hoffman, M., LaViola, J., Sottilare, R. *"Unobtrusive mood assessment for training applications*". *CHI (2011)*

APPENDIX 1 - Questionnaire

Challenge

I felt that I was learning *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I thought it was hard *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I felt stimulated *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I felt challenged *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I had to put a lot of effort into it *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I felt time pressure *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Enjoyment

I had fun *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I felt pleasure *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I felt I had enough previous knowledge to complete the task *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Enjoyment & Excitement

I enjoyed *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I felt excited *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Excitement

I felt bored *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I felt irritable *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

I found it impressive *

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree