Do We React in the Same Manner? Comparing GSR Patterns Across Scenarios

Chen Wang
CentrumWiskunde&Informatica
Science Park 123, Amsterdam the Netherlands
cw@cwi.nl

Pablo Cesar
Centrum Wiskunde&Informatica
Science Park 123, Amsterdam the Netherlands
p.s.cesar@cwi.nl

ABSTRACT
Is the physiological response from participants different between a lab experiment and a field study? In this paper, we exhaustively compare the GSR (galvanic skin response) patterns between two different scenarios. The first one was conducted in a theatre during a performance, while the second one in a laboratory during a video watching session. Questionnaires, interviews, and video recordings helped us to interpret sensor patterns, and to map them to user engagement. When comparing the GSR responses, we found a strong positive correlation between all engaged users of the two scenarios. Interestingly, such correlation was not present between the responses of non-engaged users. These results show the homogeneity of positive responses across scenarios, when compared to the variability of negative ones. The results corroborate as well that sensor data results obtained in lab studies cannot be easily generalized to real-world situations.

Author Keywords
GSR patterns; audience engagement; video consumption; theater performance;

ACM Classification Keywords
H.5.m. Information interfaces and presentation: User Interfaces – User-centered Design; J.5 Computer Applications: Arts and Humanities,

INTRODUCTION,
Physiological sensors provide valuable and reliable data about the responses of users to products and experiences. Lately, it has attracted the interest of the HCI community, becoming one more tool to help evaluations. Unlike subjective approaches like surveys, sensors provide objective data, do not interfere with the activity, and can be instrumented for different purposes (e.g., to visualize sensor data in real time).

Even though the many benefits of sensor technology for evaluating user experience, comparative studies across scenarios are missing. Most of the previous studies targeted one situation, one context, and one activity. There have been articles reporting the use of physiological sensors as objective evaluation mechanisms [16, 5]. Others have explored the relationship between biosensor data and subjective methods [15, 11]; and most of the work has been targeted to label the users' affective states [10] (e.g., fatigue). Moreover, in most of the cases the user studies were conducted in lab settings and the results were based on the averaged performance of the group (and not on the performance of the individuals).

This lack of comparative studies raises one important question then. Are user responses in controlled lab studies comparable to those obtained in the field? For example, can we use GSR data patterns reflecting non-engagement in a lab as a baseline for identifying non-engagement in the field? In general, authors discourage such generalization [6, 25], since the response of users might be different in different contextual situations. This paper compares two scenarios, and the particular case of physiological sensors, exhaustively comparing the GSR data obtained in two different studies.

In order to perform the comparisons, we need to classify user response into clusters, representing different types of feedback. We must avoid previous pitfalls of averaging data readings [13, 24], which do not provide the required level of detail and concreteness. Machine learning can play an important role for classifying data [1]. Nevertheless, we decided not to use such approach because of the annotations required in the training data. These annotations may alter the sensor readings and cannot be detected in the data set. In our experiments we prefer not to explicitly assign tasks during the experiment, which might help machine learning, but surely will disturb the experience of the users.

We classify engagement following the method initially proposed by Peter Lang [12]: GSR sensor and audience subjective reports are used together in order to identify the feedback from the users. Similarly, Celine Latulipe et al., used the same model to describe audience engagement for recorded videos of performing arts [15]. They extended the

NordiCHI '14, October 26 - 30 2014, Helsinki, Finland

ACM 978-1-4503-2542-4/14/10…$15.00.
http://dx.doi.org/10.1145/2639189.2639201

Figure 1: Two scenarios are considered: performance and video consumption. The first one (left and middle images) was studied during a theatre play, while the second one (right image) was studied in a large lab session. In both cases, sensor data was collected using exactly the same sensors and software.

model linking the GSR sensor data to two self-reported scales. Their results indicate that GSR readings are a valid approach for measuring audience engagement. Furthermore, researches in affective computing and HCI have shown interesting results between GSR and engagement [22].

In particular, we conducted an exhaustive comparison between the GSR sensor patterns across two user studies, aiming at better understanding whether engagement follows similar patterns. One experiment was run in the field, during a theatre play, and the other one was run in the lab, with users watching videos (see Figure).The same experts, using exactly the same sensors and software, ran the two experiments. In both cases, apart from the sensor data, several other materials were recorded (interviews, questionnaires, videos) in order to identify user engagement.

These materials helped us to interpret the sensory patterns, and to subjectively map them into types of audience engagement. For example, terms defined in the questionnaires were used to check audience emotional states, such as cheerful or enjoyable. Furthermore, group interviews provided us detailed information about the experience (e.g., a bad day may distract audience attention). Finally, video recordings were used to more accurately analyze the data, for example to recall what happened during the experiment and to examine, in synchronicity, particular events during the experiment and the behavior of the users.

We believe the research reported in this paper can help bridging a gap between lab and field studies, helping to better understand what to expect from physiological sensors. In particular, this paper aims to answering the following research question:

R1: Does audience engagement show correlated sensory patterns across different scenarios in a lab and the field?

This paper is structured as follows. First, we contextualize our research by discussing relevant related work. Then, we describe the experimental design and the applied methodologies, detailing the background of the participants, the data collection process, and the data analysis. Next, we report our results about classifying engagement and the comparison of the sensory data between the two scenarios. Finally, we discuss the results and elaborate on the implications of our results.

RELATED WORK

Audience engagement has been extensively studied in the past for a diverse set of domains. Christopher Peters et al. considered engagement in relationship of concepts such as perception, cognition, experience, and action [21]. Heather L. O'Brien et al. defined user engagement in online environments as the perceived usability, aesthetics, focused attention, and felt involvement [20]. Similarly, for video consumption, audience engagement has been described as the players' state of awareness and synchronization [17].

Audience engagement also has many associated and similar concepts such as audience response, audience experience, and audience feedback. Moreover, some studies used audience affective state, such as emotional states, to describe audience engagement. Peter Lang used both GSR sensor data and audience subjective reports in order to describe a two dimensional model, defining audience emotional states: valence and arousal [12]. This model has been widely applied in many later studies [15, 1, 8]. These studies employed this model as a definition of audience engagement. In particular, thanks to machine learning techniques, algorithms can even classify the four different emotional states based on the readings from one single GSR sensor [7].

Readings from GSR sensors have been used for better understanding responses to audiovisual and creative material. For example, the combination of GSR sensor readings and other biosensors have played a role for better understanding the users' affective state when playing games. These affective states have been used as an interactive component of the game [3, 18] and to assess game design strategies [9, 14, 16].

Apart from video games, GSR sensors (and a combination of a set of biosensors) plus subjective reports from users

have been applied to other applications. Some examples, run in lab settings, include identifying areas of frustration for older Web 2.0 users [14], assessing media quality [26], animating text in online chat [28], and building a real-time group interest index [17]. On the other hand, Tao Lin et al. executed a field study regarding the experience of people going to movies, combining subjective and physiological measures [17].

In addition to video consumption and gaming, GSR sensors have been used in the performing arts. Celine Latulipe et al. are key researchers in this domain, aiming at investigating the relationship between GSR responses and recorded videos of performances (lab studies). Their results demonstrate that in fact GSR readings can be used as indicator of audience engagement in these scenarios [15].

Our work extends these results (e.g., [29]) trying to better understand how different are GSR patterns in different scenarios and contexts (lab studies versus field trials). In particular, we study the results obtained from a performance (but with the audience in the theatre and not in the lab) and compare them with results obtained in the lab, where users were consuming recorded videos.

METHOD

Selection of videos

In order to select the videos participants watched, we conducted two rounds of interviews with experts: a professor at the media faculty of one of the top Chinese universities and six of his students. Based on the collected opinions, we chose two advertisements: one is a fitness product, and the other one promotes one coffee brand. Unlike previous studies about video watching, we used videos as the stimulus to capture the GSR response from users, and not to identify their emotional state [19]. This lab experiment was done in one of the computer labs the same Chinese university (Figure).

Theatre play

This experiment took place in the UK, during an interactive theatre play that lasted around 30 minutes. Four actors devised a comedy with different types of performance: juggling, asking the audience questions, and trumpet playing. Fifteen audience members participated in the play at a local theatre in the UK (Figure).

Apparatus

GSR sensors with wireless communication modules are better suited for running studies with larger groups of users. Such infrastructure should be capable of handling several sensors at the same time, since certain times it is not possible to repeat the experiment for each user (e.g., theatre play). Moreover, several rounds of repeated experiments might bring some undesirable random effects to the data collection.

In the studies we used the same home-built, with Arduino, GSR sensors (Figure 2). The wireless module was different in each study: RF12 for the video watching experiment, and Xbee for the theatre play. This implied different sample rates due to the different protocols executed at the MAC layer. The data was sampled at 1Hz using polling scheme (theatre play) and 50Hz using ALOHA (video consumption), respectively. Since we knew that ALOHA would bring more collisions, we increased the sample rate in this case. Both settings were extensively tested in the pilot studies.

The Xbee wireless module works on 2.4 GHz with relatively short communication range (roughly 10 meters). While RF12 works on 868 MHz and with a maximum of 30 meters range. In both cases, we had a sink node connected with the laptop to receive all the data packets.

All the GSR sensors were tested independently in terms of reliability and robustness before the real experiment. For example, we invited more than 50 users to watch video clips and to play video games, how our GSR sensors performed during these events. Furthermore, we also plotted the data distribution of each GSR sensor, since we know that the data patterns from GSR sensors should be a linear function.

Participants

Figure 2: Data gathering system. Our own built hardware to gather GSR data in the two scenarios

In the live theatre play, we recruited 15 users in total: seven females (Mean age = 28.29, SD = 4.85) and eight males (Mean age = 23.13, SD = 8.21). None of them had performance experience before. During the performance, the audience was required to attach the GSR sensor in their left palm. In the lab environment, the two videos were displayed one after the other to two different groups of participants. The first group, Video A, consisted of 15 users: seven females (Mean age = 22.67, SD = 3.01) and eight males (Mean age = 20.3, SD = 1.25). In the second group, Video B, 14 users took part in the experiment: seven females (Mean age = 21, SD = 2.08) and seven males (Mean age = 21.17, SD = 1.47). All the participants had GSR sensor attached in their non-dominant hand during video playback.

The culture and background of the participants in the two scenarios were rather different. In the live theatre play, all the participants were from the UK with different backgrounds. On the contrary, the participants recruited for the lab experiments were undergraduate and master students from one of the top universities in China. We agree that this might be a limitation of the comparison, since culture and background might play a role in the GSR patterns.

Questionnaires

In both cases, a pre-questionnaire and a post-questionnaire were provided before and after the experience. The majority of the questions in the post-questionnaires were related to emotions derived from either the theatre play or the videos. In the theatre experiment, questions in the pre-questionnaire were mainly about the type and intensity of the emotions they had experienced during the (working) day. In the video watching experiment, we also examined whether the participants had watched the videos before, and their previous knowledge and experience on video design. The questions were in the form of "Graphic Rating Scales" in which users were asked to make a mark on a line between two extremes, e.g.,

How often did you laugh during the performance?

Not at all Very
|_____|

The line measured 100 mm and responses were accurately measured to 1 mm.

Experimental procedure

In the play, within group design was applied. When the audience arrived, they were required to fill the questionnaires. Before the play, we explained to the audience what we were measuring during the performance, and some notes that they should pay attention to, such as not taking off the sensors during the performance. After the performance, they had the post-questionnaires.

In the video consumption, a between group design was conducted, and the experiments were run in the two rounds with the two different group participants. All of them were required to fill the pre questionnaires before the experiments. In addition to the pre questionnaires, we explained to the users about the purpose of the experiments, and some actions should be avoided during the experiments, for instance questions. When the first group finished the video, they filled the post questionnaires before they left. After that, we had the second group participants watching the second video.

Data analysis

Multidimensional scaling (MDS) is a means of visualizing the level of similarity of individual cases of a dataset, in particular to display the information contained in a distance matrix [23, 27]. Furthermore, MDS technique aims to place each object in an *N*-dimensional space such that the between-object distances are preserved as well as possible. In our analysis, we used a two –dimensional space to display the similarities between the objects.

MDS has been widely applied in psychological research [2, 4]but it is a new research technique when applied to physiological computing. Unlike other statistical techniques that test hypotheses that have been proposed a priori, MDS is an exploratory data method that explores data for which no specific hypotheses have been formed. The difference is that MDS is a means of visualizing the level of similarity of individual cases of a data set, in particular to display the information contained in a distance matrix, i.e., Euclidean distance.

	T1	T5	T7	T9	T10	T11	T12	T13	T14	T15	A1	A5	A7	A8	A9	A10	A13	A14	A15	B1	B5	B7	B12	B13
T1	1	.860**	.936**	.892**	.893**	.866**	.889**	.902**	.908**	.860**	.932**	.624**	.832**	.849**	.718**	.799**	.881**	.892**	.872**	.793**	.864**	.853**	.873**	.874**
T5		1	.950**	.917**	.842**	.928**	.944**	.871**	.925**	.827**	.731**	.756**	.939**	.753**	.710**	.788**	.861**	.822**	.945**	.947**	.913**	.703**	.736**	.919**
T7			1	.893**	.849**	.913**	.916**	.896**	.920**	.816**	.855**	.677**	.924**	.788**	.653**	.811**	.865**	.842**	.931**	.926**	.903**	.778**	.810**	.964**
T9				1	.914**	.903**	.891**	.882**	.939**	.908**	.786**	.685**	.816**	.832**	.821**	.847**	.876**	.916**	.890**	.819**	.850**	.811**	.865**	.805**
T10					1	.834**	.842**	.913**	.960**	.942**	.762**	.674**	.797**	.925**	.853**	.868**	.939**	.958**	.843**	.767**	.810**	.912**	.915**	.767**
T11						1	.891**	.911**	.893**	.813**	.752**	.692**	.856**	.753**	.667**	.748**	.819**	.808**	.878**	.859**	.811**	.762**	.796**	.862**
T12							1	.872**	.903**	.864**	.777**	.764**	.894**	.783**	.736**	.795**	.877**	.836**	.921**	.860**	.924**	.699**	.728**	.879**
T13								1	.926**	.863**	.848**	.661**	.813**	.816**	.762**	.855**	.842**	.853**	.860**	.783**	.761**	.891**	.907**	.817**
T14									1	.947**	.774**	.725**	.904**	.925**	.817**	.904**	.958**	.960**	.922**	.860**	.891**	.885**	.898**	.859**
T15										1	.719**	.684**	.820**	.960**	.907**	.894**	.965**	.978**	.861**	.695**	.865**	.880**	.872**	.755**
A1											1	.516**	.679**	.711**	.607**	.719**	.710**	.753**	.760**	.647**	.717**	.804**	.843**	.774**
A5												1	.690**	.620**	.663**	.659**	.695**	.642**	.728**	.711**	.727**	.517**	.557**	.682**
A7													1	.803**	.647**	.771**	.895**	.823**	.942**	.902**	.937**	.708**	.694**	.944**
A8														1	.819**	.849**	.965**	.976**	.811**	.662**	.836**	.909**	.874**	.739**
A9															1	.805**	.828**	.856**	.742**	.531**	.738**	.784**	.779**	.592**
A10																1	.849**	.879**	.837**	.701**	.761**	.841**	.875**	.732**
A13																	1	.971**	.889**	.784**	.920**	.847**	.827**	.836**
A14																		1	.857**	.727**	.864**	.902**	.903**	.770**
A15																			1	.854**	.925**	.759**	.786**	.922**
B1																				1	.847**	.602**	.640**	.909**
B5																					1	.697**	.689**	.915**
B7																						1	.969**	.702**
B12																							1	.705**
B13																								1

Table 1: The correlation of the responses across the red clusters: "liked the performance very much" (*: p<0.05; **: p< 0.01)

All the data analysis was done using SPSS. Pearson product-moment correlation coefficient was used to analyze the similarities or dissimilarities between the GSR readings. After that, MDS was applied to visualize the clusters of the responses on a perceptual map. In order to interpret each audience cluster, we took into consideration the subjective data for identifying the actual different types of response.

Regarding to the audience arousal level, we used the first reading coming into the sensor as the baseline for our calculation. We also investigated the arousal level in each cluster, and we displayed this result in Figure 6.

The exhaustive comparison was done by performing several times the Pearson product-moment correlation coefficient. In this way, we could examine whether the same type of responses, i.e., engaged users, was correlated with the sensory patterns across the two scenarios, and thus providing an answer to our research question (R1). Taking into consideration that the two cases (theatre play and video playback) had a different duration, we averaged the time before we performed the algorithm. This averaging procedure did not change the data distribution of sensor readings.

In the results of Pearson product-moment correlation coefficient, we used one star "*" representing 95% confidence level and two stars "**" indicating 99% confidence level. In addition, the overall fit statistics (Kruskal's stress and R Square) in MDS were provided to reveal how the algorithm fitted the input data.

All the GSR sensor data were post-processed using the smoothing and filtering Matlab function, in order to minimize the impact of hand movements. However, we found that thanks to the well-prepared design of the experiments, the data were of high quality. Overall, there was no much difference in the data before and the data after the smooth procedure.

RESULTS

Audience clustering

In Figures 3, 4 and 5, we display the MDS results from the two user studies. We used T plus a numerical numbers (sensor id), e.g., T3, to represent the participants from the theatre play, and either A (the video A: fitness product) and B (the video B: coffee brand) accompanied with the numerical numbers (sensor id) to represent to a user who joined the lab experiments. We used four different colors (red, yellow, green and orange) to distinguish the different clusters, where the same color in all the figures represents the same category of responses.

Figure 3 displays the four audience clusters in the theatre play, showing the different experiences. Users in the red clusters reported the highest scores in the post questionnaires in terms of cheerful, enjoy and like. On the other hand, participants in the green clusters scored the lowest in the questionnaires.

In addition to the questionnaires, subject T3 in the interview told us that he did not like the play at all, and this can be seen in the recorded video. Users in the yellow and orange clusters had a different experience: T2 and T8 took a while to get into the performance, which might imply that they did not understand the beginning of the performance. In contrast, T4 and T6 attention waned in half through the

performance, one of the users in the interview stated that she started to recognize one of the actors at certain moment of the play and that is why her attention shifted.

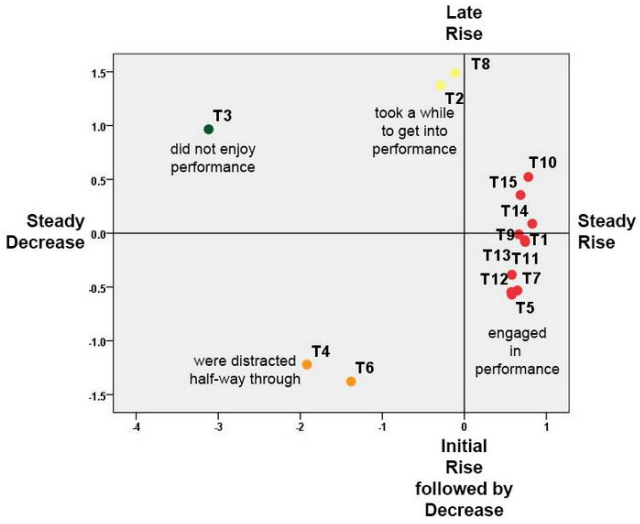

Figure 3: Feedback clusters for the theatre play (Stress: 0.03, RSQ: 0.99)

Figure 4 displays the clusters from the responses of users watching video A. In this particular case, we only observed three types of feedback. Similarly than to the previous figure, the red cluster represents users that rated the highest scores in terms of immersion, attention level, and concept design (encouraging them to purchase the product). By contrast, participant A11 (the green dot) was not so interested in buying the product, and he labeled his attention level as the lowest possible. Interestingly, the participants in the yellow cluster reported that they had no previous experience, and their knowledge was limited on the video design.

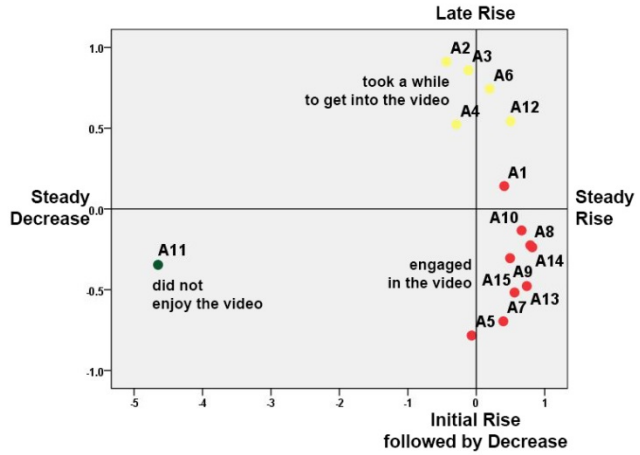

Figure 4: Feedback clusters for Video A (Stress: 0.03, RSQ: 0.99)

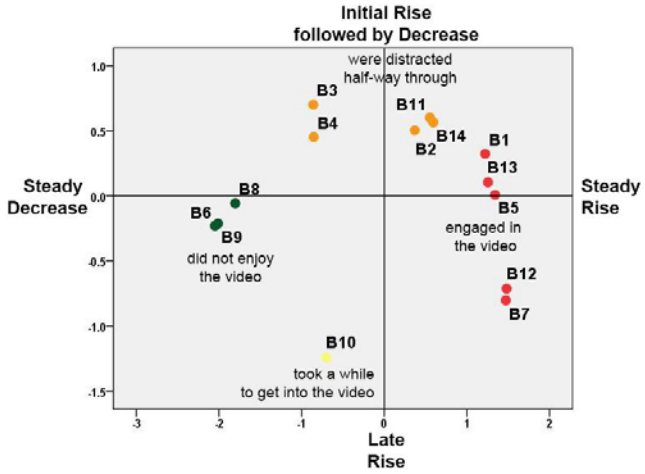

Figure 5: Feedback clusters for Video B (Stress: 0.04, RSQ: 0.99)

In Figure 5, we find four clusters. Similarly, the participants in the red cluster rated the highest score in terms of immersion, and attention level; the participant B10 (the yellow point) reported that he had no previous experience and knowledge on video design. During the group interview, the students from the orange clusters were rather active, and show interest in this video, but they all reported a busy day and this might explain why their attention declined after a while after the video started playing.

Figure 6 shows the arousal level in each cluster from the two user studies. Obviously, the theatre play evoked a higher arousal compared to consuming video. On the other hand, the participants from both the red cluster and the orange one all had positive arousal levels, which were much higher than the ones from the rest of the clusters: the arousal levels of the green clusters were all negative values, and the yellow clusters had relevant low arousal scales, displaying a negative value for video B.

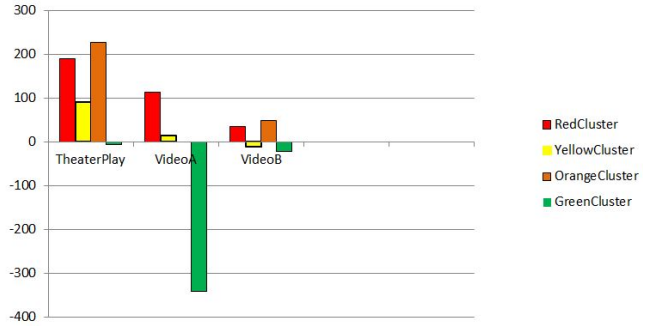

Figure 6: Arousal levels from the clusters in each experiment

By observing the distribution of data in the four clusters in both studies, we found that there were roughly four different patterns: a steady increase in the red clusters; a steady decrease in the green clusters; a late rise in the yellow clusters; and initial rise followed by a decrease in the orange clusters. These descriptions were used to label the figures generated by MDS.

Through the combination of the GSR feedback and the subjective data, we could clearly identify two main types of responses: the engaged audience versus the non-engaged audience, which were in the red clusters and the green clusters respectively. However, with respect to the users in the yellow and the orange clusters, we concluded that personal reasons interfered their watching experience.

Response Patterns (R1)

We can assume that participants in the same color cluster had a similar watching experience. Our interest was on examining whether their GSR sensory patterns showed correlations across the two scenarios.

Regarding the red cluster, we were surprised to see that all the users' GSR responses (24 users) were all significantly correlated with each other (Table 1): averaging 0.831. We can safely conclude that the sensory patterns were strongly synchronized across the scenarios.

For the orange and yellow clusters, the GSR responses were partially correlated (Table 2 and Table 3). For instance, the GSR response of participant T4 was strongly correlated to all the users that attended the video experiment, user T6 was correlated to most of the users in the orange clusters. Nevertheless, both T2 and T8 were both correlated to most of the users located in the yellow clusters of the video consumption.

For the green clusters on Table 4, we found that the correlation was not strong. Correlations only existed between the two watching video experiences: averaging 0.872, p< .01, and there was no significant correlation between video consumption and theatre performance. In addition to the correlation checking, we also found that the GSR response of T3 experienced fluctuations during the performance, although all the users displayed a steady decrease on their sensory pattern.

	B2	B3	B4	B11	B14	T4	T6
B2	1	.336	.233	.689**	.496**	.701**	.267
B3		1	.706**	.553**	.495**	.675**	.426*
B4			1	.294	.518**	.398*	.290
B11				1	.657**	.929**	.703**
B14					1	.690**	.597**
T4						1	.683**
T6							1

Table 2: The correlation of the responses across the orange clusters: "got distracted during the performance" (*: p<0.05; **: p< 0.01)

	T2	T8	B10	A2	A3	A4	A6	A12
T2	1	.871**	-.311	.444*	.502**	.278	.644**	.737**
T8		1	-.091	.629**	.740**	.362*	.871**	.884**
B10			1	.502**	.328	.247	.226	.126
A2				1	.854**	.635**	.810**	.722**
A3					1	.586**	.894**	.849**
A4						1	.610**	.616**
A6							1	.937**
A12								1

Table 3: The correlation of the responses across the yellow clusters: "took a while to understand the performance" (*: p<0.05; **: p< 0.01)

	T3	A11	B6	B8	B9
T3	1	-.107	.004	-.066	.076
A11		1	.900**	.761**	.815**
B6			1	.894**	.968**
B8				1	.894**
B9					1

Table 4: The correlation of the responses across the green clusters: "did not like the performance" (*: p<0.05; **: p< 0.01)

We performed MDS on the GSR responses of the non-engaged participants (Figure 7) and the GSR response of all them (Figure 8), in order to investigate the distance, in a perceptual map, between the non-engaged participants and the proximity between the non-engaged ones and the rest of the them.

In Figure 7, we found that the distance was large between the responses of T3 and the responses of the other users. This result is consistent with the results displayed in the Table 4.

In Figure 8, we could clearly see a massive cluster formed in the left part of the map, mainly coming from the red, orange and yellow clusters. In contrast, the responses from the non-engaged video consumers formed a cluster on the right side of the map, where the green points were closed to each other. Regarding the responses from subject T3, his geometrical location was more adjacent to the orange clusters, and positively correlated to the responses from B2: 0.549, p< 0.01; negatively correlated to the ones from B10, A2 and A3: averaging -0.501, p< 0.01.

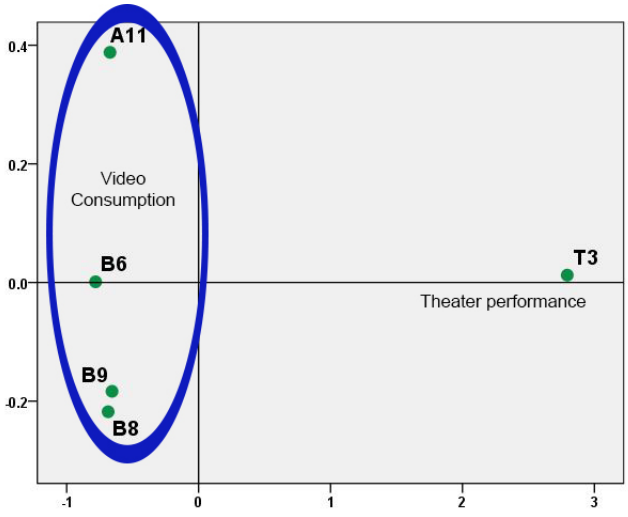

Figure 7: MDS result when applied to all the responses in the green clusters (Stress: 0.05, RSQ: 0.99)

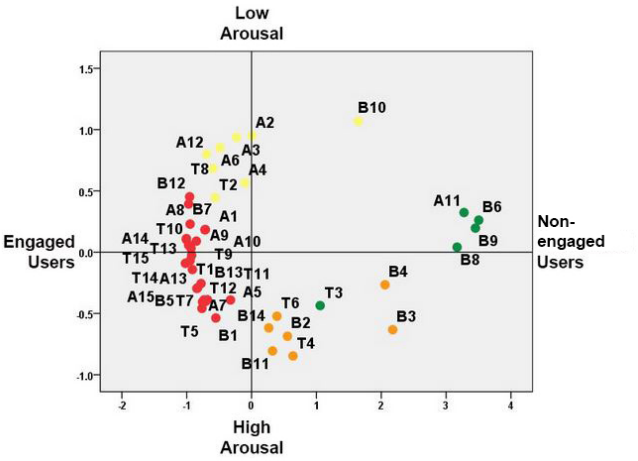

Figure 8: MDS result when applied to all the responses across the two scenarios (Stress: 0.03, RSQ: 0.99)

DISCUSSION

In this paper, we have compared the GSR responses from participants in two different use cases. In particular, we used MDS to successfully classify audience responses in clusters and subjective data (interviews, questionnaires) to interpret what each cluster represents. Based on these techniques, we could differentiate between engaged and non-engaged participants, and then perform exhaustive comparison across the two use cases.

We found that the responses from the engaged users showed a strong correlation on their sensory patterns between lab and field studies. Interestingly, the responses from the non-engaged participants did not correlate across use cases (between lab and field trial), but correlated between the two lab experiments. This result is consistent with a similar phenomenon mentioned in previous research [6, 25]: a "boredom" state captured in a lab may have the different patterns compared to the one in a field study. Still, these previous studies did not quantify such results and did not report any comparative data for the more engaged users.

Even though we could use "boredom" to define the state of the participants in the green clusters, we decided to apply the more general term, non-engaged, to define this type of responses. We followed the learning from our previous studies: it is unlikely to generate a boredom state when people are watching short videos in a lab situation, since every participant might take the task seriously. They typically try to understand what it is happening in the video, ignoring its quality even if the video is in an unknown language. On the other hand, "boredom" is a state that can instead happen in longer field trials. Therefore, we preferred to refer the responses in the green clusters as non-engaged, which better describes the cases for the lab situation and the field trial.

In our studies, we only had 5 out of 44 users with non-engaged responses. We certainly may require larger amount of this type of responses to better understand how similar are sensory patterns across a lab study and a field trial. Nevertheless, it is almost impossible to estimate the number of non-engaged users that would result from an experiment, even though a large number of participants (44 in our case) are involved.

It would be a strong backup to our findings if we can provide some results of the subjective evaluations, for example whether the questionnaires reports of the engaged users were also strongly correlated. However, in our case we had some difficulties to run the correlations across the two scenarios. First, the questions designed in the two use cases were not exactly the same ones, so that it is not possible to correlate the answers from one participant joined the lab experiment to the one attended the theater play. Second, the correlation method requires the sufficient and equal length of data sets to run the algorithm, but in questionnaires each participant gave one score for each question, which is unlike the sensor data: each participant had a sequence of sensor readings (from the beginning of the theater play/video consumption to the end). Therefore, the subjective data crossing the two scenarios cannot satisfy these requirements, but we can compare the scores of engaged users to the non-engaged ones. Besides that, the video recordings and the interviews also helped us interpret the clustering results of the MDS.

We believe that comparing a video - recorded performance and a live performance is another interesting research topic. In our case, we had the different experimental settings: the theater play versus the video consumption. However, we think that the results are valid and innovative. First, the sensory patterns of engaged users are strongly correlated crossing the different scenarios, even though the experimental settings are different, and this phenomenon has never been mentioned in the previous studies. Second, more interestingly, the sensory patterns of the non-engaged users are different, which motivates us to take a further step to investigate the reasons that may cause this result. For instance, whether the experimental settings have effects on the non-engaged users, or the matters of the performance itself, as we have already seen the sensory patterns of the non-engaged users were strongly correlated in the lab experiment as shown on Figures 7, 8. At the current stage, we will leave this box open for the future exploration.

The methodology we have applied for reporting the results does not require intentional inputs from the participants, in order to guarantee high quality of the sensor data. However, it is still a challenge to interpret sensor data, which requires well-designed experiments, questionnaires, proper-organized interviews, and high quality of video recordings.

CONCLUSION

In this paper, our exhaustive comparison between GSR responses has shown a strong correlation between the sensory patterns of the engaged users across a lab study and a field trial. With respect to the non-engaged users, we

could not make such concrete conclusion. Even though similarities existed across the two scenarios, there was no strong correlation.

The results imply that engaged users display a significant similarity on their GSR response across the two scenarios. Based on this finding, we can potentially label GSR patterns as engaged, instead of having to use audience annotations. Moreover, in the future we may get rid of the subjective data, and identify engagement based only on the GSR responses. The benefits therefore are twofold: we can avoid the impact, on the experiment, of constant labeling the responses of the users, and for the sensory patterns of engaged users we do not need to use the more subjective techniques (questionnaires, interviews, video recordings).

On the other hand, the results reported in this article require further research on non-engagement. In our comparison we found the two rather different sensory patterns from the lab to the field. In order to better understand the representative GSR patters for non-engagement might require many other user studies trying to better understand the contextual impact, the previous experience of the participants, and the profile. This paper takes a first step in this direction, which we believe it is useful, and we hope other researchers will follow up and help us to provide a more complete picture about the behavior of physiological sensors in different situations.

However, our solution has some limitations, which will not be able to identify the specific affective state of the users (e.g., fatigue) since only one sensor was used. Therefore, we believe that extra sensors will certainly provide us more rich sensory data (and patterns) regarding audience engagement, In terms of techniques we believe MDS provides a fantastic tool for mapping sensory patterns with types of responses, still we would like to explore if and how machine learning can be applied.

REFERENCES

1. Mohammad Adibuzzaman, Niharika Jain, Nicholas Steinhafel, Munir Haque, Ferdaus Ahmed, Sheikh Ahamed, and Richard Love. 2013. In situ affect detection in mobile devices: a multimodal approach for advertisement using social network. *SIGAPP Appl. Comput. Rev.* 13, 4 (December 2013), 67-77. DOI=10.1145/2577554.2577562
2. Ingwer Borg and Patrick J.F. Groenen. Modern Multidimensional Scaling: Theory and Applications. 2005 *Springer Science+Buiness Media, Inc.* ISBN-10: 0-387-25150-2.
3. Guillaume Chanel, Cyril Rebetez, Mireille Trancourt, and Thierry Pun. 2008. Boredom, engagement and anxiety as indicators for adaptation to difficulty in games. In *Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era* (MindTrek '08). ACM, New York, NY, USA, 13-17. DOI=10.1145/1457199.1457203
4. Trevor F. Cox and M.A.A. Cox. *Multidimensional Scaling, Second Edition.* ISBN 1-58488-094-5.
5. Joel E. Fischer and Steve Benford. 2009. Inferring player engagement in a pervasive experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '09). ACM, New York, NY, USA, 1903-1906. DOI=10.1145/1518701.1518993
6. Stephen H. Fairclough. Fundamentals of physiological computing. *Interact. Comput. (2009) 21 (1-2): 133-145.*
7. Rui Guo, Shuangjiang Li, Li He, Wei Gao, Hairong Qi, and Gina Owens. 2013. Pervasive and unobtrusive emotion sensing for human mental health. In *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare* (PervasiveHealth '13). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, 436-439. DOI=10.4108/icst.pervasivehealth.2013.252133
8. Christian Martyn Jones and Tommy Troen. 2007. Biometric valence and arousal recognition. In *Proceedings of the 19th Australasian conference on Computer-Human Interaction: Entertaining User Interfaces* (OZCHI '07). ACM, New York, NY, USA, 191-194. DOI=10.1145/1324892.1324929
9. Kai Kuikkaniemi, Toni Laitinen, Marko Turpeinen, Timo Saari, Ilkka Kosunen, and Niklas Ravaja. 2010. The influence of implicit and explicit biofeedback in first-person shooter games. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10). ACM, New York, NY, USA, 859-868. DOI=10.1145/1753326.1753453
10. Ahish Kapoor, Windlow Burleson and Rosaline W. Picard. Automatic prediction of frustration. Int. J. *Human-Computer Studies (2007)*
11. Tao Lin, Akinobu Maejima, and Shigeo Morishima. 2008. Using subjective and physiological measures to evaluate audience-participating movie experience. *In Proceedings of the working conference on Advanced visual interfaces* (AVI '08). ACM, New York, NY, USA, 49-56. DOI=10.1145/1385569.1385580
12. Peter J. Lang. The Emotion Probe: Studies of Motivation and Attention. *American Psychologist, May 1995.*
13. Henry Ledgard, Designing for the usability. *Human Aspects of Computing. Communications of the ACM*, March 1985, Volume 28.

14. Darren Lunn and Simon Harper. 2010. Using galvanic skin response measures to identify areas of frustration for older web 2.0 users. *In Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)* (W4A '10). ACM, New York, NY, USA, , Article 34 , 10 pages. DOI=10.1145/1805986.180603
15. Celine Latulipe, Erin A. Carroll, and Danielle Lottridge. 2011. Love, hate, arousal and engagement: exploring audience responses to performing arts. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11). ACM, New York, NY, USA, 1845-1854. DOI=10.1145/1978942.1979210
16. Pejman Mirza-Babaei, Lennart E. Nacke, John Gregory, Nick Collins, and Geraldine Fitzpatrick. 2013. How does it play better?: exploring user testing and biometric storyboards in games user research. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13). ACM, New York, NY, USA, 1499-1508. DOI=10.1145/2470654.246620
17. Anmol Madan, Ron Caneel, and Alex "Sandy" Pentland. 2004. GroupMedia: distributed multi-modal interfaces. *In Proceedings of the 6th international conference on Multimodal interfaces* (ICMI '04). ACM, New York, NY, USA, 309-316. DOI=10.1145/1027933.1027983R. Mandryk. Objectively evaluating entertainment technology. *In CHI'04, pages 1057–1058*. ACM Press, 2003
18. Lennart Erik Nacke, Michael Kalyn, Calvin Lough, and Regan Lee Mandryk. 2011. Biofeedback game design: using direct and indirect physiological control to enhance game interaction. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11). ACM, New York, NY, USA, 103-112. DOI=10.1145/1978942.1978958
19. Eva Oliveira, Mitchel Benovoy, Nuno Ribeiro, Teresa Chambe. Towards Emotional Interaction: Using movies to automatically learn users' emotional states. *INTER ACT 2011, part I*, lncs 6946. Pp. 152-161,2011.
20. Heather L. O'Brien and Karon E. Maclen Measuring the User Engagement Process. *Engagement by Design Preconference Workshop*, CHI 2009 Digital Life New World, Boston, MA, April 5, 2009.
21. Christopher Peters, Ginevra Castellano, Sara de Freitas. An exploration of user engagement in HCI. *AFFINE '09, November 6,*2009. Boston,MA, USA.
22. R.W.Picard. Affective computing. MIT Press, Cambridge, MA, USA,1997.Shengsheng Ruan, Ling Chen, Jie Sun, and Gencai Chen. 2009. Study on the change of physiological signals during playing body-controlled games. *In Proceedings of the International Conference on Advances in Computer Enterntainment Technology* (ACE '09). ACM, New York, NY, USA, 349-352. DOI=10.1145/1690388.1690456
23. Schiffman, Susan S., M. Lance Reynolds, and Forrest W. Young (1981), *Introduction to Multidimensional Scaling: Theory, Methods, and Applications*, NY: Academic Press.
24. M. S. Sridhar, Understanding the user- why, what and how? *Library Science with a slant to Documentation and Information Studies*, 32 (4), December 1995, 151- 164.
25. Hari Sundaram. 2013. Experiential media systems. *ACM Trans. Multimedia Comput. Commun. Appl.* 9, 1s, Article 49 (October 2013), 4 pages.
26. Gillian M. Wilson and M. Angela Sasse. 2000. Do users always know what's good for them? Utilizing physiological response to assess media quality. *Proceedings of HCI 2000: People and computers XIV – Usability OR ELSE*.
27. Young, Forrest W., and Robert M. Hamer (ed.) (1987), *Multidimensional Scaling: History, Theory, and Applications*, Hillsdale, NJ: Erlbaum.
28. Hua Wang, Helmut Prendinger, and Takeo Igarashi. 2004. Communicating emotions in online chat using physiological sensors and animated text. *In CHI '04 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '04). ACM, New York, NY, USA, 1171-1174. DOI=10.1145/985921.986016
29. Chen Wang, Erik N. Geelhoed, Phil P. Stenton, and Pablo Cesar. 2014. Sensing a live audience. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (CHI '14). ACM, New York, NY, USA, 1909-1912. DOI=10.1145/2556288.2557154