Quantifying Emotional Responses of Viewers based on Physiological Signals

Sriram Raju Dandu
Electrical and Computer Engineering
University of Virginia
Charlottesville, USA
sd9aj@virginia.edu

Geoff Gill
Shimmer Americas
Boston, USA
ggill@shimmersensing.com

Caleb Siefert
Deparment of Behavioral Sciences
University of Michigan
Dearborn, USA
csiefert@umich.edu

***Abstract*—This paper presents a novel metric that extracts features from ECG and EDA to quantify the responses of each viewer as none, medium, and high in near real time. This metric is intuitive and corresponds well with manual coding of biometric data. Initial evidence for the validity and reliability of this metric comes from a series of studies using diverse stimuli (i.e. sporting events; advertisements; situational comedy) and a diverse set of participants (N = 231). Implications for the usefulness of this approach for a variety of purposes are discussed.**

Keywords—Media research, physiological signals, heart rate, skin conductance, inflection point, signal processing, adaptive thresholding.

I. Introduction

An emotional response can be defined broadly as reactions to a particular intrapsychic feeling or feelings, accompanied by physiological changes that may or may not be outwardly manifested. In short, emotions involve physiological arousal and cognitive attribution (i.e., labeling). Emotional arousal involves changes in the activity of the autonomic nervous system. The autonomic nervous system (ANS) is a general-purpose physiological system responsible for modulating peripheral functions, such as heart-rate and respiratory pace [1]. This system consists of sympathetic and parasympathetic branches, which are generally associated with activation and relaxation, respectively. Because of the general-purpose nature of the ANS, its activity is not exclusively a function of emotional responding, but rather encompasses a wide variety of other functions related to digestion, homeostasis, effort, attention, and so forth.

Commonly assessed indices of ANS activation are based on electrodermal (i.e., sweat gland) or cardiovascular (i.e., blood circulatory system) responses. Electrodermal activity, is the property of the human body that causes continuous variation in the electrical characteristics of the skin. Measurement of EDA, commonly known as skin conductance (SC), is modulated autonomously by sympathetic activity which drives human behavior, cognitive and emotional states on a subconscious level. SC offers direct insights into autonomous emotional regulation. Najström and Jansson [2] investigated the predictive value of skin conductance reactivity in response to masked threatening pictures on emotional responding following stressful life events. EDA can be used to index several different processes such as activation, attention, and the task significance or affective intensity of a stimulus [3]. Simons et al. [4] showed that SC responses were greater for high-arousal, compared to low-arousal, images. Hubert and de Jong-Meyer [5] also found that a 10-min cartoon film induced a pleasant amused state characterized by low levels of arousal and a rapid decrease in SC. By contrast, a suspense film induced a reduction in relaxation, and an increase in arousal exhibited with a marked increase in SC.

The most commonly used cardiovascular measures for studying emotional responses include heart rate (HR), blood pressure (BP), total peripheral resistance (TPR), cardiac output (CO), pre-ejection period (PEP), and heart rate variability (HRV). HR and BP reflect a combination of sympathetic and parasympathetic activity, and HRV has been closely linked to parasympathetic activity [6]. Choi et al. [7] suggested a HRV-based evaluation only when a high level of emotion is induced by visual stimulation. De Jonckheere et al. [8] showed that the HRV based metric could be a good indicator of parasympathetic changes in emotional situation. Appelhans and Luecken [9] provided a theoretical and empirical rationale for the use of HRV as an index of individual differences in regulated emotional responding. HRV is an accessible research tool that can increase the understanding of emotion in social and psychopathological processes.

Researchers have integrated ANS outputs into studies focusing on emotional responses to various stimuli and media. Detenber et al. [10] investigated the effects of picture motion on individuals' emotional reactions to images with the help of subjective measures and physiological data (SC and HR). Renaud and Blondin [11] measured heart rate, frequency of skin conductance responses, and self-reports to study the anxiety levels of participants during performance of a computer version of the Stroop Color–Word Interference Test.

As an advertising medium, television offers numerous advantages, such as audiovisual strength, great geographic coverage, superior target selections, and mass audience reach [12]. To secure these merits, considerable financial investment is necessary. For instance, the average price for a 30-second commercial in the 2018 Super Bowl was over $5 million [13]. Thus, to maximize these benefits and to minimize financial waste, advertising practitioners and researchers have paid attention to the effectiveness of television advertising. Several articles have shown that effectiveness of advertisements can be studied by monitoring the neurophysiological parameters. Siefert et al, [20] collected biometric data during SuperBowl 2008 and studied the utility of biometric measures for assessing consumers' emotional engagement with advertising content. Ohme et al. [14] demonstrated that neurophysiological measures can capture differences in consumer reactions to slightly different marketing stimuli. The authors have shown significant differences in neurophysiological reactions to an altered scene in a TV commercial. Ravaja [15] provided an overview of the use of psychophysiological measures of attention and emotion in media research with the focus on 3 most commonly used measures: heart rate, facial electromyography, and electrodermal activity. Lang [16] analyzed heart rate data to show short-term attentional responses and longer-term arousal in subjects viewing commercial messages.

There is little debate about the potential benefit of evaluating ANS response to stimuli. However, the cost, time, and difficulty of interpretation of such studies has severely limited the utilization of these techniques. The latter can be particularly vexing, as many who could benefit from inclusion of physiological measures refrain from employing them due to lack of familiarity, experience, or perception that they will be unable to interpret data outputs. This paper presents a novel metric for studying emotional responses of subjects based on their heart rate and skin conductance that addresses these issues. This metric can be generated in real time, requires no baselining, is intuitive and easy to interpret, and can be used to evaluate different stimuli on the same scale.

This paper describes the calculation of the metric, the system used to collect the data and presents a series of studies that demonstrate the validity and reliability of this metric. The primary contribution is the demonstration of a system which can collect biophysical data, compute non-conscious metrics and generate results with no human intervention in real time.

II. Pilot Data Collection

A. Participants

Data was collected from 74 subjects including males and females in a number of sessions. The subjects comprised of college students and adults. The participants were seated in a closed room and the stimuli were projected or displayed on a screen. The participants were briefed about the sensors and the motivation of the study before the advertisements started playing.

B. Apparatus

All the data studies were performed using NeurolynQ developed by Shimmer. Participants wore the sensor on their wrist. A total of four electrodes were connected to the sensor. Two electrodes were placed across the heart to capture the ECG signal. Two electrodes were placed on index and middle fingers to measure the EDA. All the electrodes were connected to the sensor for streaming to the desktop in real-time and for local storage. The sensor calculates inter-beat intervals (IBI) and SC in real time from the ECG and EDA data using standard signal processing techniques. The IBI and SC values were transmitted at 5Hz to the desktop for displaying the live measurements and for analyzing the patterns to identify emotional responses in real time.

C. Stimulus Material

Several months after Super Bowl, participants were shown four Super Bowl advertisements in a closed room. There was a 5 second countdown clock shown to participants between the advertisements. The four Super Bowl advertisements were Kia Niro, Mr. Clean, World of Tanks and KFC. As per several consumer reports (AdAge, Boston Globe Media Partners, Adweek, etc.), Kia Niro and Mr. Clean had high positive ratings whereas World of Tanks and KFC had low ratings. This paper presents an algorithm to compute non-conscious metrics for these ads and later compares them with the ratings.

III. Methodology

This section explains the techniques implemented to extract features from an individual's IBI and SC data. It also describes the encoding technique for quantifying emotional response based on the derived features at an individual level.

A. SC Response

Three types of SC patterns that are of interest are shown in the Figure 1. These three patterns are labelled as SC responses, namely, "High Positive Trend", "Positive Trend" and "Inflection Point". The definitions of the SC responses are mentioned in Table 1.

Table 1: Definitions of SC Responses

SC Response	Definition
High Positive Trend	The rate of increase of SC over time is high based on a threshold value.
Positive Trend	The rate of increase of SC over time is non-negative.
Inflection Point	The rate of increase of SC over time shifts from negative to positive.

Ravaja [15] provided an overview of the use of skin conductance and trend-based metrics in media research. Researchers [24, 25] presented methods to identify inflection points in skin conductance.

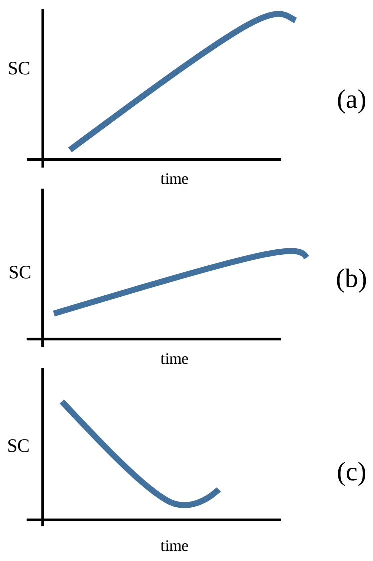

Figure 1: Skin conductance patterns (a) shows high positive trend, (b) shows positive trend and (c) shows inflection point.

All the three SC responses are dependent on the slope of the values within the window. A linear regression model is incorporated to calculate the slope of SC values within the window. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. In our case, the two variables are time and SC values respectively. Mathematically, a linear regression can be represented as:

$$Y = \mathbf{a}X + c \tag{1}$$

where Y are SC values, x is time, c is a constant and **a** is the slope.

$$\mathbf{a} = (X'X)^{-1}X^{T}Y \tag{2}$$

In simple words, linear regression model identifies the rate at which the SC values are increasing or decreasing during the window interval.

If the slope is greater than a positive threshold, then it is labelled as “High Positive Trend”. The threshold is dependent on the window, thus making it adaptive and participant centric. Figure 2 shows the SC values of two participants watching the same advertisement. The slope of participant 2 is lower than that of participant 1, but slope of participant 2 is relatively high when compared with neighboring values. This is coherent with the claim that all the participants respond in a different way as the physiology of a human body differs from individual to individual. One participant might be aroused to a particular stimulus whereas another might not show any interest. Thus, an adaptive thresholding scheme is chosen to build an unbiased segregation criterion.

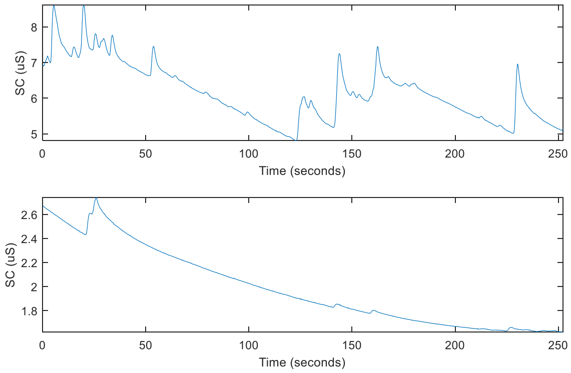

Figure 2: The plots of skin conductance of two participants when exposed to the same stimuli. Top Plot: Participant 1; Bottom Plot: Participant 2.

The adaptive threshold is a factor of average of the SC values within the window. If the slope is above this threshold, then the window is labelled as “High Positive Trend”. If the slope is lower than the threshold but is non-negative, then the window is labelled as “Positive Trend”.

The “Inflection Point” metric is based on slope as well. This metric identifies the points of a curve at which a change in the direction of curvature occurs. If an individual’s overall slope of SC is negative in a window but shows sign of an uptick, then the window is marked as “Inflection Point”. This metric facilitates in early detection of SC response to a stimulus compared to trend or peak detection schemes as these schemes can only detect responses after the event has occurred. Methods presented in [24, 25] identify inflection points in post-hoc processing. For our purpose, the following method was implemented to detect these points in real-time if the last SC value is above a certain threshold over the negative trend line, then it is labelled as “Inflection Point”. Figure 3 shows the SC responses of the participants shown in Figure 2.

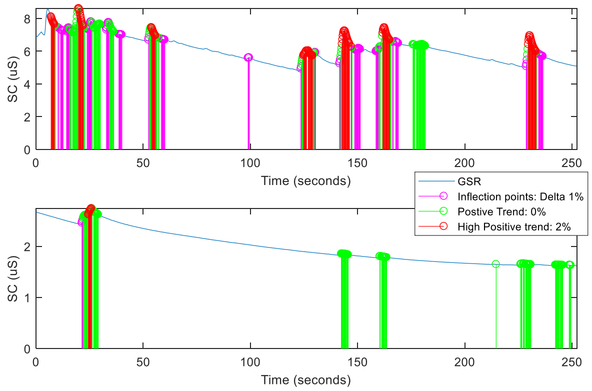

Figure 3: SC responses of two participants when exposed to the same stimuli. The blue lines are the SC values, green markers are for positive trends, red markers are for high positive trends and magenta markers are for inflection points.

B. HRV Response

Yamaguchi et al. [20] mentioned that emotional events rapidly and automatically capture attention. Specifically, stimuli that are more emotional in nature attract more attention relative to stimuli that are less emotional in nature. So, it's of little surprise that indicators of attention, such as HRV, are often employed in research on emotion [21, 22]. HRV has been shown to be an effective means of assessing the attention levels of participants. There is extensive research on the use of HR for identifying emotion responses when participants are exposed to a stimulus [7]–[9]. If the HRV is deviating from normal range for a subject, it indicates that the participant is responding to a stimulus.

The HRV metric is generally computed using HR or IBI data collected over long time intervals i.e. a few hours or minutes. Malik et al. [23] suggested a 5-minute epoch size for calculating HRV using time-based or frequency-based methods. Esco et al. [17] showed there is no significant difference between HRV based on 1-minute and 5-minute epochs. Goss and Miller [18] implemented EBC (Estimated Breadth Cycle) to obtain HRV using 10-second epochs. Authors showed high correlation between EBC and other established HRV metrics. So, we incorporated the EBC metric to obtain HRV values in real time. Similar to SC responses, a HRV response is detected if the EBC is not within a certain range. Research shows that while resting the HR varies slightly and is not constant, thus HRV at resting is non-zero. If the HR is constant for an interval of time (low HRV), it is considered that the person is paying attention or engaged in the task, whereas if the HR varies more than a resting person (high HRV) then the participant was aroused.

So, if the HRV is close to zero or above a certain threshold, it is labelled as an HRV response to the stimuli. Since this variation of HR is dependent on the individual, an adaptive thresholding approach is implemented like that of SC response. The range is dependent on the mean IBI interval within the window, making the thresholds unbiased and participant centric. There is a higher and a lower HRV thresholds, and if the HRV is not within these limits, the window is marked as a HRV response as shown in Figure 4.

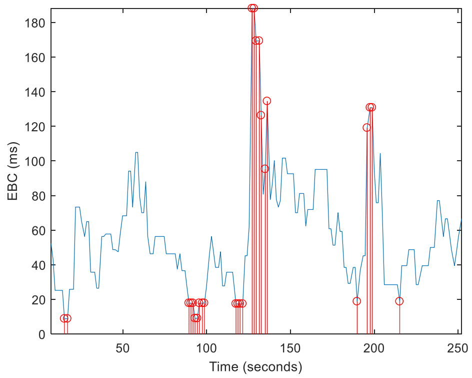

Figure 4: Plot of HRV values of a participant. Blue – EBC values and Red – HRV Response.

C. Response Categorization

The individual's emotional responses are coded based on the derived SC and HRV responses. The responses are segregated into categories, namely, 'High Response' and 'Response'.

Table 2: Response Categorization Matrix

SC Response	HRV Response	No HRV Response
High Positive Trend	High Response	Response
Positive Trend	Response	Response
Inflection Point	High Response	Response
No SC Response	No Response	No Response

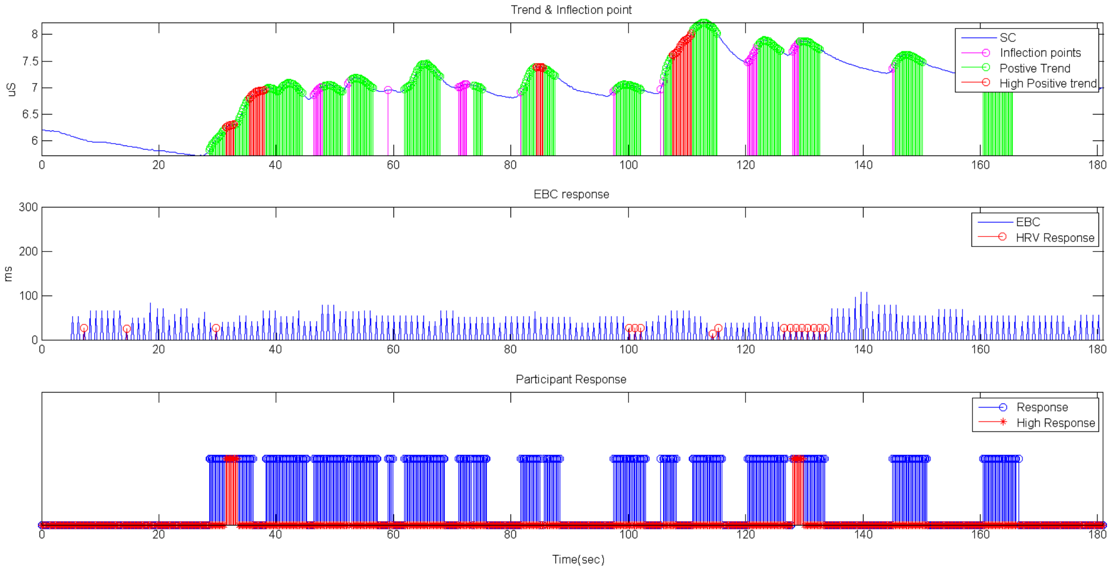

Figure 5: The responses of a participant while watching the Super Bowl advertisements. The yellow, green, orange and black lines near time axis show the time intervals of Kia, Mr. Clean, World of Tanks and KFC advertisements respectively.

For our purposes, ‘High Responses’ are labelled when SC starts to increase monotonically and a HRV change occurs simultaneously. Our decision model marks high responses when *either* a “High positive trend” or “Inflection point” occurs in conjunction with a simultaneous HRV response. Similarly, ‘Responses’ are labelled when SC increases slightly from its baseline. Our decision model marks responses when either a “Positive trend” or “Inflection point” occurs. Figure 5 shows the quantified emotional responses of a participant when he/she saw the four Super Bowl advertisements.

IV. Pilot Observations

The derived SC and HRV responses are coded in a subtle but simple way to quantify the emotional responses. The SC responses consists of three levels: high positive trend, positive trend and inflection point. The HRV response is a binary metric i.e. based on the thresholded EBC value.

An example of trend and inflection points for SC are marked in the Figure 5. The magenta points identify the onset of the positive slope; the green points mark the positive trend and the red points show the high positive trend of SC. There is no sharp deviation of HRV for most of the time. There were a couple of intervals when the HRV was low, which means that the heart rate remained almost constant and suggests the participant was interested or paying attention.

The participant responded to the stimuli most of the time. Near 30 seconds, there is a high positive SC trend and HRV response, so it is labelled as high response. Similarly, near 130 seconds, an inflection point and HRV response occur at the same, which is also a high response. The rest of the “responses” were due entirely to SC (meaning there was an absence of corresponding HRV response). From Figure 5, it is evident that the participant showed responses throughout the first 2 advertisements but did not respond much to the latter two advertisements (in-line with the consumer report, mentioned earlier).

Researchers are generally not inclined to study a single individual in most cases. Thus, responses from all participants can be aggregated by simply calculating the percentage of the audience showing a response over time. As you can see in Figure 6, near 35 seconds 100% of the audience is showing a “response” for the Kia advertisement, 75% for the Mr. Clean advertisement, 37.5% for World of Tanks advertisement and 62.5% for KFC. The percentage of audience who showed high response is more significant in the former two advertisements relative to the latter two. The percentage audience plot will allow moderators to observe the instances when most of the participants are engaged. All the above metrics can be calculated in real-time setup, as the algorithms to extract the responses and categorize them rely on the data collected in a 5-second window and not on the entire data from the study. This equips moderators of focus groups to identify points of interests in advertisements or other stimuli in real-time rather than depending on the self-reports which are attained after the advertisement. The self-reports provide an opinion of the entire advertisement or the stimuli but does not pin point to the particular event which might have been of interest to the viewer. Self-reports might be biased due to external factors which are independent of the stimuli, whereas the algorithm identifies the changes in the physiological parameters which are not affected by external parameters.

V. Validation Studies

A. Introduction

The primary purpose of the pilot study was to develop a largely automated method for coding audience member responses into three levels (i.e., no response; a response; a high response). Subsequent studies were then undertaken to examine the

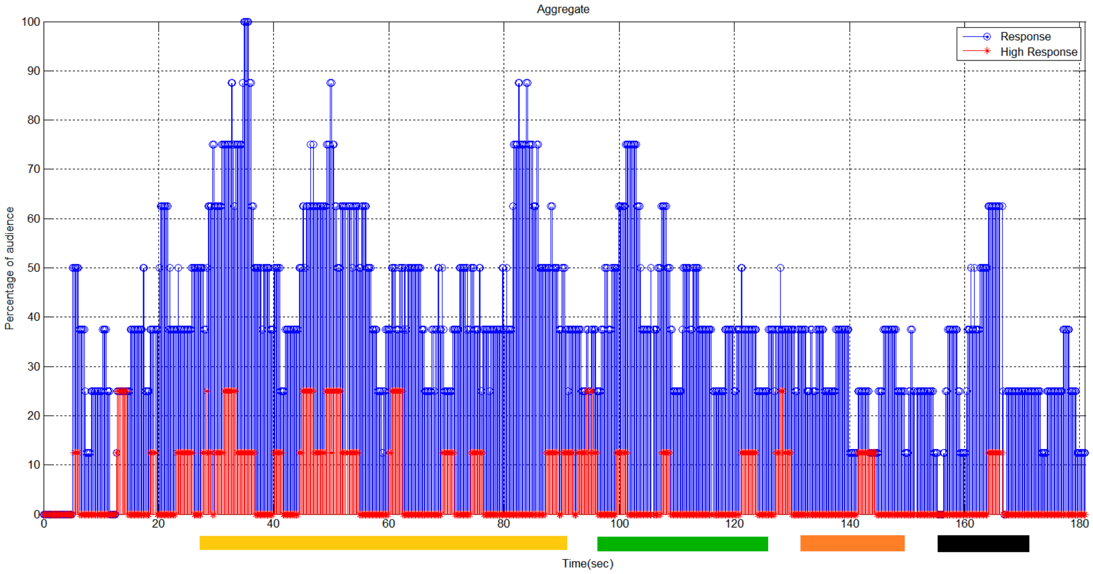

Figure 6: The percentage of audience who responded over the course of the advertisements. The yellow, green, orange and black lines near time axis show the time intervals of Kia, Mr. Clean, World of Tanks and KFC advertisements respectively.

validity of the metrics. We were particularly interested in determining if the metric was truly measuring emotional arousal. Finding an indisputable outcome measure of emotional arousal is challenging. In the pilot study, we did find that advertisements that were rated more positively were more likely to produce emotional arousal responses, as indicated by our metric, in a larger number of audience members. But, as has been well demonstrated in other literature [15], non-conscious measures provide information on emotion that is complimentary, or even contradictory, to that provided by self-report.

To address the challenge of finding an independent outcome measure, we decided to evaluate the response metrics during sporting events. There is little question that some plays in a sports contest are more emotionally arousing than other plays. For example, we would expect scoring plays to be more emotionally arousing relative to the average non-scoring play. While some non-scoring plays may be exciting, most are likely to be less emotionally arousing relative to scoring plays. Thus, if our metric showed greater arousal for scoring plays relative to non-scoring plays, it would provide some support that the response metric is assessing emotional arousal. Additionally, we expected that scoring plays would be more arousing than standard emotional media content (e.g., a sitcom).

The data we used for this analysis was a combination of four different studies, all with broad samples of males and females, where we collected GSR and HR on a total of 154 people:

1. SuperBowl 2018: 45 subjects watching the SuperBowl 2018 game live (Advertisements and halftime show included). The subjects were evenly split between fans of the two teams and the data was collected in New York.
2. Fifa World Cup (WC) 2018: 42 fans of soccer watching the France vs Peru game live (Halftime ads included). Note: The data was collected in London and only one of the participants had a strong affiliation to either team.
3. FIFA World Cup (WC) 2018: 16 subjects watching the Mexico vs Brazil game live (Halftime ads included). The data was collected in Mexico and all subjects were fans of Mexico.
4. A 20-minute Situation Comedy (Sitcom): a total 51 subjects were monitored in seven different sessions while they watched an unaired episode of a TV show. The study was conducted in Las Vegas and demographics of the participants was very broad with the primary requirement that they watch sitcoms. We include the sitcom because we will use this data to evaluate reliability and as another data point.

Table 3: List of Stimuli

SuperBowl	**Fifa WC** *France-Peru*	**Fifa WC** *Mexico-Brazil*	**Sitcom**
Game	Game		TV Show
Ads	Halftime Ads		
Field Goals	Goals		
Touchdowns			

For all above data recordings, participants were allowed to move freely and react naturally. The only instructions about how to consume the content were given in the sporting events where participants were asked to pay attention to the advertisements (which were the primary target of the studies).

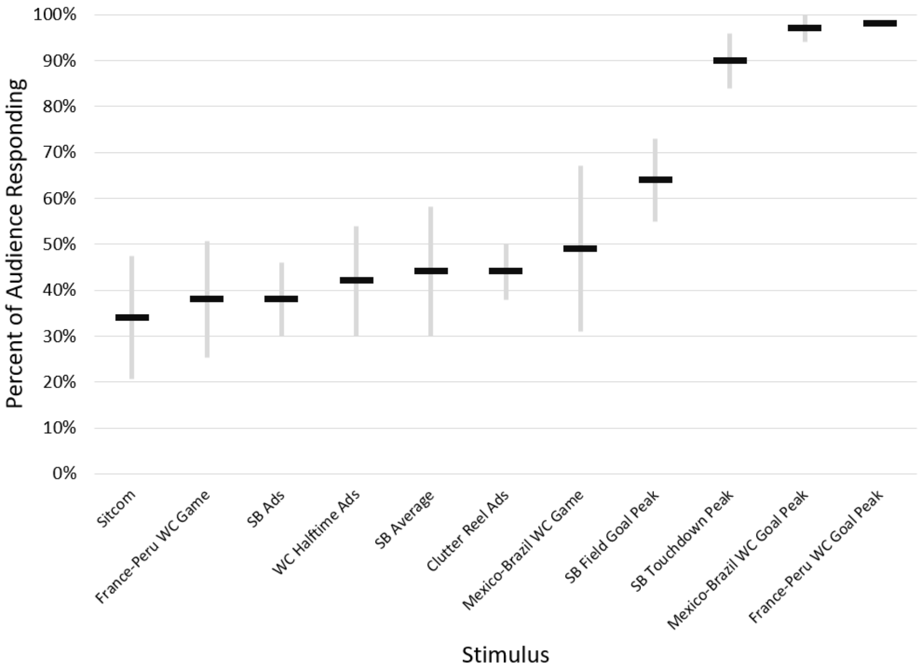

Figure 7: Response levels to different stimuli.
SB and WC stand for Super Bowl and World Cup respectively.

‘Any’ responses for all the datasets were computed with the algorithm described previously. For the rest of the paper, ‘Any’ responses are referred to as responses.

The different stimuli the next section will compare are mentioned in Table 3.

B. Discussion

Figure 7 summarizes the results of the response data for all the different stimuli mentioned in Table 3. The horizontal bars represent the average level of the specific stimuli and the vertical bars represent the standard deviation of the sample (there was only one goal in the France-Peru WC game, so there was no variance to calculate).

We developed six hypotheses for our metrics that could be tested with these data sets. Each is discussed below, ranked by the size of the expected effect:

1. **Soccer goals should be higher than the average for the game.** It is intuitively obvious that the scoring of a goal should provoke more response than the average in a soccer game. In both the Fifa WC games peak percent of responding audience to goals is more than four standard deviations higher than the average response during game. The chance that three random samples would be more than three sigma above the mean is about one in 400 million.
2. **Super Bowl touchdowns should be higher than the average play.** Similarly, football touchdowns are more arousing than the average play. We monitored 143 plays in SuperBowl 2018. On average 44% of the audience responded (with a standard deviation of 14% while watching these plays which included TDs, FGs and other non-scoring plays. Out of these 143 plays, we monitored seven touchdowns, during which the peak response was 90% of the audience, 3.2 standard deviations above the average for the game.
3. **Super Bowl field goals should be lower than touchdowns, but higher than the average play.** This is because field goals are for fewer points, they are set up in advance (so they are not a surprise), and the outcome is generally predictable, based on the location in the field from which the kick is taken. Still, they are a scoring play and therefore more important that the average play. With field goals, the average peak response rate was 64%, 1.4 standard deviations higher than the average for the game, but well below the average for touchdowns, exactly as expected. The chance of this happening randomly is less than 0.1%.
4. **The average for advertisements should be lower than for the games themselves.** 44% of audience responded on average during the Super Bowl 2018 game, but only 38% responded to the advertisements. In the France-Peru WC game, 38% of the audience responded to the game but only 30% responded to the halftime advertisements. In the Mexico-Brazil WC game, 49% of audience responded during the game and 48% to the advertisements. In all cases, the number of viewers who respond to a game is slightly higher compared to the advertisements, supporting the hypothesis. The differences are not as great as one might expect; however, the instructions to pay attention to the ads may have inflated the advertisement response.
5. **The average level of response during the game should be lower for the France-Peru game than for the Mexico-Brazil game.** We expect this response because the France-Peru game did not have fans of either team, whereas the Mexico-Brazil game had fans of Mexico. In fact, there was a significant difference, of 38% responding on average to France-Peru, compared to 49% for Mexico-Brazil. This difference may have been increased by the fact that the France-Peru game may have been a less interesting game on average (a number of participants indicated afterwards that it was a boring game).
6. **The average levels of these high-profile games are likely to be higher than that of a Situation Comedy.** Based on viewership and the inherent nature of competition, one would expect high profile sporting events to be more exciting than a sitcom – especially one with poor viewership that was cancelled soon after the study. The data supports this hypothesis, in that the average percent of viewers who responded to the show is less compared to all the three athletic events. The average percentage of audience responding to a sitcom was 34% whereas for SB, France-Peru WC game, Mexico-Brazil game is 44%, 38% and 49% respectively.

The analysis of these four datasets supported all six of the a priori hypotheses. These results support the contention that the response metric:

1. Captures moments which cause emotional arousal in viewers.
2. Differentiates stimuli which cause emotional arousal to those stimuli which don’t.
3. Quantifies different levels of emotional arousal among viewers.
4. Compares different stimuli on the same scale.

VI. Reliability

A. Introduction

51 people in total were monitored during a standard ~20 minute TV sitcom using NeurolynQ to quantify the reliability of the algorithm and determine sample size requirements in a realistic study context. A total of seven different sessions were recorded in a standard focus group facility. Between 4 and 10 people participated in each session as outlined in the Table 3. Participants were recruited by mall intercept with broad

demographics: general population, male and female, between 18 and 60 years old.

Table 4: Size of Individual Sessions

Session #	Number of participants
1	8
2	7
3	8
4	10
5	4
6	6
7	8
Total Population	51

B. Analysis

When determining sample size, the key criterion is “whether one would make similar decisions based on subset of population as one would if it were based on the total population.” The “Any Response” metric calculates the percentage of audience who showed a either a medium or high response in a time interval. A peak in the Any Response trace means that higher number of participants are responding to a particular stimulus. These peaks are referred to as “Key moments”. The question is whether these Key Moments are repeatable in different samples responding to the same stimulus and how large a sample size is needed to see the Key Moments consistently.

The authors quantified the definition of a Key Moment as a peak in the Any Response trace which is above a certain threshold i.e.

$$Threshold = Average + \tau \times Standard\ Deviation \qquad ()$$

where τ is threshold factor.

Researchers use “Key moments” to decide whether a stimulus has made a substantial impact on the audience. So, if a key moment exists in both the subset and total population around the same time interval, then one can interpret the same conclusion with smaller sample as they would with a larger sample size. If that is the case, the results are reliable. Determining a threshold factor (τ) for defining a key moment may vary depending on the application. Furthermore, there should be some tolerance around the exact size of the key moment because the decision making should be the same whether the moment is 0.9 or 1.1 sigma above average. This is particularly important as sample sizes decrease. For example, we evaluated sample sizes down to four people, where the limited number of possible outcomes (0%, 25%, 50%, 75%, and 100%) make matching the exact size of the key moment unlikely. Therefore, we chose a threshold factor (τ) of 1 in identifying key moments in the control sample and 0.5 when determining if they were replicated in the test sample (as indicated below).

The following three metrics are employed to compare the performance of the different sample sizes:

$$Hit\ Rate = \frac{Nboth}{Ntotal}$$

$$Miss\ Rate = \frac{Nmiss}{Ntotal}$$

$$False\ Discovery\ Rate = \frac{Nfalse}{Nsub}$$

where *Nboth* is the number of key moments that exist in both the total population and the subsample around the same time interval, *Ntotal* is the number of key moments in the total population, *Nmiss* is the number of key moments that exist in in the total population but not in the subsample, *Nfalse* is the number of key moments that existing in the subsample but not in the total population and *Nsub* is the number of key moments in the subsample.

Hit rate represents the degree of similarity whereas miss rate and false discovery rate shows the degree of dissimilarity. For hit rate and miss rate, the threshold factor was 1.0 for the total population and 0.5 in the subsample. For false discovery rate, it was 1.0 in the subsample and 0.5 in the total population.

C. Sample Size and Reliability

Any Responses for all seven individual sessions and different combinations of sessions { (6,7), (2,4), (1,4), (1,3,5), (3,5,7), (1,3,6), (1,6,7), (1,2,3), (1,2,4), (3,5,6,7), (1,2,3,4,5) } were computed. The combinations were chosen to gather different population sizes between 14 and 37. Figure 8 and Figure 9 display the hit rate and false discovery rates for the different population sizes when compared with the total population. Trendlines show that hit rate approaches 1 and false discovery rate approaches 0 as the population size increases. Mathematically, $Miss\ Rate = 1 - Hit\ Rate$, so we will not plot it.

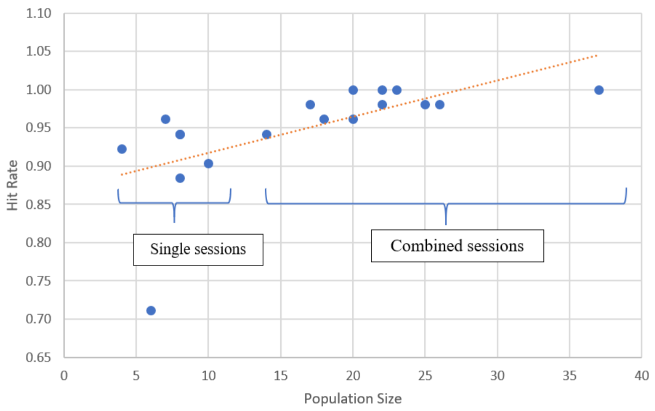

Figure 8: Hit rates for different population sizes.

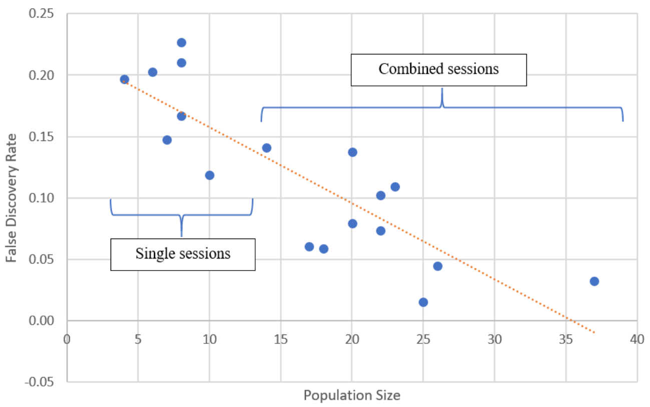

Figure 9: False discovery rates for different population sizes

For individual sessions, average hit rate is about 90%, which means 1 out of 10 key moments in the total population are absent in sample size of <=10. When 2-3 sessions are combined, average hit rate is 98%, whereas when 5 sessions are combined, the hit rate is 100%. Similarly, average false discovery rate for individual sessions is 18%, for 2-3 combined sessions it is 9% and for 5 sessions combined it is 3%. This means that individual sessions contain a lot of false key moments which aren't present in the total population. Hit rate increased and false discovery rate dropped when sessions are combined compared to individual sessions.

Data from individual sessions will be more likely be affected by extraneous events (e.g., someone coughing, a loud laugh, etc.). All participants might be affected such events, creating a response for that that group alone. Such extraneous key moments would not be repeated across sessions.

To study the influence of extraneous events on repeatability, 20 participants were randomly picked from across all seven sessions and the analysis was repeated. Only 3% of the key moments in the subset data were not reflected in total sample, whereas it is 9% when a couple of sessions were combined. This shows that aggregation across several sessions reduces false discovery rate by 67%, thereby producing repeatable results by reducing the impact of extraneous events.

Table 5 summarizes the results by showing the averages for the metrics for the individual sessions and combined sessions that total between 17 and 23 people (average 20.3 people).

Table 5: Performance Metrics of Different Sample Sizes

Sample size	Hit rate	Miss Rate	False Discovery Rate
4-10 (Single Sessions)	90%	10%	18%
17-23 (Combined two sessions)	98%	2%	9%
20 (All sessions)	98%	2%	3%

The false discovery rate (< 10%) and hit rate (~98%) show that in order to obtain fewer false positives, at least a sample size of 20 is required. To validate the above statement, a t-test was performed between sample sizes of less than 10 and around 20,

$$\text{t-statistic} = -2.74 \quad \& \quad \text{p-value} < 0.05$$

Thus, the difference in hit rate between a sample size of 20 and the individual sessions is statistically significant. Finally, we evaluated "Low Periods" (defined as periods where the response rate was less than one sigma *below* average) in all the subsamples (both individual sessions and combined samples). This represents a test of the frequency of false negatives. We found that there was NEVER a Key Moment ($\tau = 1$) during a Low Period. Thus, if there is a Low Period in a subsample, it can be concluded with a high degree of certainty that there would not be a Key Moment in that period, if the sample were expanded.

VII. Summary and Conclusion

This paper proposes a signal processing approach using physiological signals (ECG and EDA) to quantify the emotional responses of subjects (the Response Rate metric). The SC data is categorized into one of the 3 categories: high positive trend, positive trend and inflection point. The IBI data is converted to HRV, which is further labelled as HRV responses. The SC and HRV responses are then translated to emotional responses using a unique and simple coding scheme. The response rate metric has a set of highly desirable functional features, including that it:

1. Can be calculated in real time.
2. Is intuitive to researchers and the contributions can be traced to individual respondents.
3. Corresponds well with manual coding of the data.
4. Captures the entire period of response, not just the peak.
5. Can be used to compare different stimuli on the same scale.

By integrating the hardware, software, and algorithm into a single system, researchers are able analyze emotional responses of people in real time when exposed to stimuli.

To assess validity, the authors analyzed sporting events where it could be reasonably assumed that spectators would have a high response to scoring plays. We then extended this analysis to other stimuli including advertisements and a TV situation comedy. The findings supported intuitive assumptions:

1. High excitement scoring plays (touchdowns and goals) were >3 standard deviations above average in response.
2. Lower excitement scoring plays (field goals) were ~1.4 standard deviations about average.
3. A lower involvement audience had lower response rates than a high involvement audience.
4. TV advertisements and the situation comedy had lower average response rates than the sporting events.

Finally, the authors evaluated the reliability of the metrics by comparing subsamples with the full sample in the situation comedy study. The metrics demonstrated good reliability characteristics:

1. Running 3+ sessions with a total of ~20 people will capture virtually all (~98%) of the key moments that would be captured with a sample 2-3 times as large.
2. The false positive rate on ~20-person subsamples was higher (~9%), likely because of extraneous stimuli in the sessions. If eliminating false positives is critical, a slightly higher sample size, or more sessions, may be required.
3. The false negative rate was zero for all subsamples, including those as low as four participants.

In these studies, the response rate metric demonstrated good ability to detect responses reliably, even in the lowest involvement case (of the situation comedy). The key challenge in study design is to eliminate responses to extraneous stimuli that are not part of the study question. That challenge exists for any true measure of non-conscious response, especially one that is utilized in natural (relatively uncontrolled) surroundings.

VIII. Acknowledgement

The authors would like to thank Ipsos Neuro and Behavioral Sciences Center, GfK, National Research Group, Interactive Video Productions, Dr. Joanne Dugan, Shimmer Research Inc. and participants for their support in this effort.

References

[1] I. B. Mauss and M. D. Robinson, “Measures of emotion: A review,” Cogn. Emot., vol. 23, no. 2, pp. 209–237, Feb. 2009.

[2] M. Najström and B. Jansson, “Skin conductance responses as predictor of emotional responses to stressful life events,” Behav. Res. Ther., vol. 45, no. 10, pp. 2456–2463, Oct. 2007.

[3] M. E. Dawson, Schell Anne M., and D. L. Filion, “The electrodermal system.,” in Handbook of psychophysiology, Second., Cambridge University Press, pp. 200–223.

[4] R. F. Simons, B. H. Detenber, T. M. Roedema, and J. E. Reiss, “Emotion processing in three systems: the medium and the message,” Psychophysiology, vol. 36, no. 5, pp. 619–627, Sep. 1999.

[5] W. Hubert and R. de Jong-Meyer, “Autonomic, neuroendocrine, and subjective responses to emotion-inducing film stimuli,” Int. J. Psychophysiol. Off. J. Int. Organ. Psychophysiol., vol. 11, no. 2, pp. 131–140, Aug. 1991.

[6] J. Cacioppo, G. Berntson, J. Larsen, K. M Poehlmann, and T. A Ito, “The Psychophysiology of Emotion,” in The Handbook of Emotion, 2000, pp. 173–191.

[7] K.-H. Choi, J. Kim, O. S. Kwon, M. J. Kim, Y. H. Ryu, and J.-E. Park, “Is heart rate variability (HRV) an adequate tool for evaluating human emotions? – A focus on the use of the International Affective Picture System (IAPS),” Psychiatry Res., vol. 251, no. Supplement C, pp. 192–196, May 2017.

[8] J. De Jonckheere, D. Rommel, J. L. Nandrino, M. Jeanne, and R. Logier, “Heart rate variability analysis as an index of emotion regulation processes: interest of the Analgesia Nociception Index (ANI),” Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf., vol. 2012, pp. 3432–3435, 2012.

[9] B. M. Appelhans and L. J. Luecken, “Heart rate variability as an index of regulated emotional responding,” Rev. Gen. Psychol., vol. 10, no. 3, pp. 229–240, 2006.

[10] B. H. Detenber, R. F. Simons, and G. G. B. Jr, “Roll ‘em!: The effects of picture motion on emotional responses,” J. Broadcast. Electron. Media, vol. 42, no. 1, pp. 113–127, Jan. 1998.

[11] P. Renaud and J.-P. Blondin, “The stress of Stroop performance: physiological and emotional responses to color–word interference, task pacing, and pacing speed,” Int. J. Psychophysiol., vol. 27, no. 2, pp. 87–97, Sep. 1997.

[12] J. Z. Sissors and R. B. Baron, Advertising Media Planning, 6th ed. Chicago,IL: McGraw-Hill.

[13] Carroll, Charlotte, “Super Bowl LII: How Much Does a Commercial Cost,” *Sports Illustrated*, January 11, 2018.

[14] R. Ohme, D. Reykowska, D. Wiener, and A. Choromanska, “Analysis of neurophysiological reactions to advertising stimuli by means of EEG and galvanic skin response measures.,” J. Neurosci. Psychol. Econ., vol. 2, pp. 21–31.

[15] N. Ravaja, “Contributions of Psychophysiology to Media Research: Review and Recommendations,” Media Psychol., vol. 6, no. 2, pp. 193–235, May 2004.

[16] A. Lang, “Involuntary Attention and Physiological Arousal Evoked by Structural Features and Emotional Content in TV Commercials,” Commun. Res., vol. 17, no. 3, pp. 275–99, 1990.

[17] M. R. Esco, H. N. Williford, A. A. Flatt, T. J. Freeborn, and F. Y. Nakamura, “Ultra-shortened time-domain HRV parameters at rest and following exercise in athletes: an alternative to frequency computation of sympathovagal balance,” Eur. J. Appl. Physiol., pp. 1–10, Nov. 2017.

[18] C. F. Goss and E. B. Miller, “Dynamic Metrics of Heart Rate Variability,” ArXiv13086018 Q-Bio, Aug. 2013.

[19] Siefert, Caleb J., et al. "Winning the super “buzz” bowl: how biometrically-based emotional engagement correlates with online views and comments for super bowl advertisements." *Journal of Advertising Research* 49.3 (2009): 293-303.

[20] S. Yamaguchi and K. Onoda, “Interaction between Emotion and Attention Systems,” *Front Neurosci*, vol. 6, Sep. 2012.

[21] C. Chen *et al.*, “Detecting Sustained Attention during Cognitive Work Using Heart Rate Variability,” in *2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2010, pp. 372–375.

[22] K. Laumann, T. Gärling, and K. M. Stormark, “Selective attention and heart rate responses to natural and urban environments,” *Journal of Environmental Psychology*, vol. 23, no. 2, pp. 125–134, Jun. 2003.

[23] M. Malik *et al.*, “Heart rate variability: Standards of measurement, physiological interpretation, and clinical use,” *Eur Heart J*, vol. 17, no. 3, pp. 354–381, Mar. 1996.

[24] A. De Clercq, B. Verschuere, P. De Vlieger, and G. Crombez, “Psychophysiological Analysis (PSPHA): A modular script-based program for analyzing psychophysiological data,” *Behavior Research Methods*, vol. 38, no. 3, pp. 504–510, Aug. 2006.

[25] S. R. Green, P. A. Kragel, M. E. Fecteau, and K. S. LaBar, “Development and validation of an unsupervised scoring system (Autonomate) for skin conductance response analysis,” *Int J Psychophysiol*, vol. 91, no. 3, pp. 186–193, Mar. 2014.