34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Consistency and Validity of Self-reporting Scores in Stress Measurement Surveys

Khalid Masood, Beena Ahmed, *Member IEEE,* Jongyong Choi, *Student Member, IEEE* and Ricardo Gutierrez-Osuna, *Senior Member, IEEE*

Abstract—**Stress has been attributed to physiological and psychological demands that exceed the natural regulatory capacity of a person. Chronic stress is not only a catalyst for diseases such as hypertension, diabetes, insomnia but may also lead to social problems such as marriage breakups, suicide and violence. Objective assessment of stress is difficult so self-reports are commonly used to indicate the severity of stress. However, empirical information on the validity of self-reports is limited. The present study investigated the authenticity and validity of different self-report surveys. An analysis, based on a three-pronged strategy, was performed on these surveys. It was concluded that although subjects are prone to systematic error in reporting, self-reports can provide a useful substitute for data modeling specifically in stress evaluation where other objective assessments such as determination of stress using only physiological response are difficult.**

I. Introduction

Stress is a form of physical strain that arises from the reaction of the body to outside challenges, either physical or psychological. It induces a transition that takes an individual from a calm state to an excited one through a series of physiological reactions [1], [2]. Short-term stress or an instantaneous reaction is beneficial in dangerous scenarios, but chronic stress may lead to diseases such as hypertension, insomnia, diabetes, asthma and depression [3], [4]. Long-term stress may also contribute to social problems such as marriage breakups, family fights, road rage, suicide and violence [5], [6]. Hence there is a need to monitor and report the patient's personal state for long periods of time to find stress-related symptoms.

The response to stress is a combination of physiological adaptations to reestablish the balance that was disturbed due to a stressor. Blood pressure, electrodermal activity, heart rate, respiration, pupil dilation and electromyography can be used as measures for the monitoring of stress [7]. Inter-heartbeat intervals (heart rate variability) are used as reliable measures for the determination of stress: increases in stress lead to lower heart rate variability [8], [9]. Electrodermal activity (EDA) is also widely used for the measurement of stress: increase in stress also increases skin conductivity due to sweating and perspiration [10]. Respiratory signals also provide suitable measures of stress. During stress, respiratory rate increases and breathing patterns become irregular [11].

Due to individual variability among subjects, the physiological impact of stress is different on each individual based on his physical and mental strength. Hence any inference drawn from collected physiological data needs to be correlated to perceived measures or severity of stress, typically obtained using self-reporting scores (SRS) [12]. The alternative option of objective methods is not available in our case as determination of stress from only physiological response is very challenging. Although self-reports may provide information regarding outcomes of the events, it is recommended that they do not be used as a sole measure [13]. A number of studies have reported the potential inaccuracy of self-reporting measures [14]. Generally, there are two types of inaccuracy in self reports: failure to recall the exact details, and social desirability effects. Problems with the recall are related to the subjects' inability to remember the exact details of the events, though these errors can be minimized by carefully designing the surveys to better aid in memory recall. Social desirability causes subjects to tailor their answers to present their behavior as being more favorable and less questionable. In this way, either willingly or unconsciously, reports maybe slightly distorted.

In the present study, we conducted experiments to address the following questions about the accuracy and validity of self-reporting measures of mental stress.

- Is self-reporting data consistent? Are self-reporting errors biased?
- Which survey type has the least reporting errors? Can physiological data be mapped onto perceived stress levels?

In these experiments, physiological data was collected from subjects while they performed a range of stressful and relaxing activities for which they also provided perceived stress levels. In section II, we provide a brief introduction of the hardware platform used, physical activities and methods for feature extraction. Section III describes the three-pronged strategy used to evaluate the performance of these surveys. In sections IV and V, classification model and results are discussed. Section V presents our conclusions from this study and some important future directions.

II. Methods

A total of 19 subjects (10 male, 9 female) participated in the study, for which prior approval had been obtained from the Texas A&M University Institutional Review Board. The subjects were briefed on the experimental procedure and their written consent to participate was obtained prior to commencement of the experiments.

A. Physical activities

In the experiment, subjects were asked to self-score a range of activities on a 7-point Likert scale (1 corresponds to relaxed conditions such as deep breathing and 7 corresponds to highly stressful situations such as color word test). To determine their perceived stress levels, subjects were asked

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

to complete three surveys: stress, difficulty and post-stress surveys. In the stress and difficulty surveys, the activity was ranked immediately after its completion in terms of (1) perceived stress induced by the activity and (2) perceived difficulty of the task. Subjects were asked to rank both-stress and difficulty as perceived stress depends only in part on the absolute level of task difficulty. In the post-stress survey, the subjects ranked each of the individual activities after completing the whole range of activities in the experiment. The short time delay introduced in the post-survey and exposure to subsequent activities provides opportunity to the subjects for re-assessment of the activities.

At the start of experiments, we used deep breathing to help the subjects relax: we also used deep breathing between stressful activities to help relieve the effect of the previous stressor. During deep breathing, subjects were asked to breathe deeply at a pace of 0.1 Hz for three minutes. The stress challenges consisted of a dual tracking task (subjects had to track a moving target in a computer screen using the mouse and click a mouse button whenever a target letter appeared on the screen), memory search (subjects had to memorize a set of words and then identify them among various confounders), mirror tracing (on paper printout, a pattern had to be traced manually by looking through a mirror), Stroop test (after being shown one of four words displayed in different ink colors, subjects had to click on one of the four buttons according to the ink color), supine and tilt (subjects take supine and tilted positions) and public speech (subjects have to deliver a 4-minute public speech).

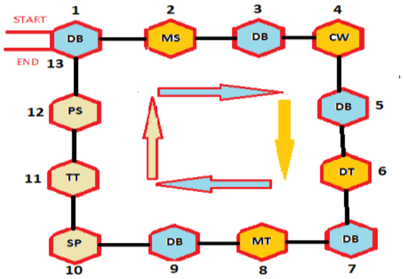

Figure 1. Block diagram of Activities: Deep Breathing (DB), Memory Search (MS), Color Word (CW), Dual Task (DT), Mirror Trace (MT), Supine (SP), Tilt (TT) and Public Speech (PS) represent mental challenges.

B. Wearable sensor system

The wearable sensor platform contained a heart rate monitor (HRM), a respiratory sensor and an EDA sensor [15]. Heart rate variability was calculated from a HRM (Polar Electro Inc.) widely used in fitness monitoring. The respiration sensor (SA9311M; Thought Technology Ltd.), measures variation in pressure exerted by the rib cage, which are then used to determine expansion and contraction changes in the lungs during respiration. Changes in skin conductivity are monitored by placing two electrodes on the proximal phalange of the index and middle finger using E243 electrode (Vivo Metric Systems Corp.).

C. Feature extraction

We extracted 13 features from the physiological signals of 19 subjects, six HRV features, 3 features from respiratory spectra and four features from EDA. Features were calculated using 90 sec windows with an overlap of 80 sec. The features include HRV power inhigh and low frequency range, mean and standard deviation of successive R-R intervals, the portion of RR interval that changes more than 15 msec (pNN15) and the root mean square of successive differences of R-R. We also computed the low and high frequency power in the respiratory spectra and their ratio. The skin conductivity signal consists of two components, skin conductance level (SCL) and skin conductance response (SCR). To extract features from EDA, a regularized least-squares detrending method is used, where the aperiodic trend corresponds to the SCL and the residual is assumed to correspond to SCR [16]. The mean and standard deviation of these components were taken as EDA features.

III. Comparative Analysis

The validity, relevancy and authenticity of a self-reporting survey can be determined by comparing subjective scores provided by a user against those provided by the user population at large. The variance in scores was used to check the consistency of a survey. The relevancy of the surveys was measured by the correlation between the scores and the corresponding physiological features. Finally, the accuracy of the surveys was determined by computing the deviation terms in each survey, where a subject was considered to have made a deviation if his score and physical parameters were inconsistent with those of other subjects.

A. Consistency

For each activity (only initial deep breathing activity is selected so in total eight activities are included in this analysis), we computed the variance across all subjects for the subjective SRS to determine the consistency of the surveys. First, variance across subjects for each activity was computed. Then an overall variance was determined as the average value of all the computed variances. The lower the variance in the scores, the more consistent the survey was and vice versa. Figure 2 shows the computed variance for each activity. The difficulty survey had the lowest variance, followed by the stress scores, whereas the post-stress survey had the highest variance. The overall variance values for each survey are presented in Table I.

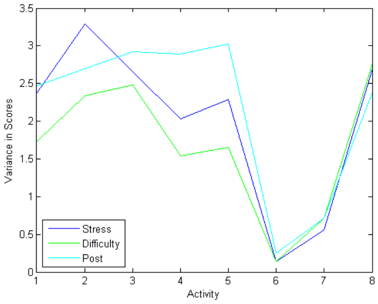

Figure 2. Variance in each activity (eight activities are selected as only one deep breathing is included) for the scores of three surveys

B. Relevancy

The correlation between the survey scores and physical parameters was calculated to determine the relevancy of the self-reported scores. Instead of using all of the available physiological parameters, EDA components are used to determine deviations. For deep breathing activities, respiration rate was controlled and HRV is trivially related to respiration. Using EDA ensured that the deviations were identified with a measure which was least affected by respiration. We used the first principal component of EDA (first principal component to represent all four features of EDA). The correlations of the first principal component of EDA with the self-reporting scores obtained from the three surveys are presented in Table I. Higher correlation values indicate that the variation in the physical parameter is similar to that of the corresponding survey. As seen, the stress survey has a higher correlation with EDA than the difficulty and post-stress surveys.

TABLE I. PERFORMANCE MEASURES FOR SURVEYS

	Stress	*Difficulty*	*Post stress*
Variance	2.01	1.67	2.17
Correlation	0.39	0.33	0.28
Error	6.9	7.2	10.5

C. Accuracy

We used a threshold based model to compute the deviation in the scores obtained from the three surveys. We compared the three scores provided by the subjects (S_i) and a representative physiological parameters (F_i) to the corresponding median values (M_s and M_f) obtained from all the subjects for each activity i to identify discrepancies. The values which fell outside a range centered on the median were marked as having discrepancies. This median range was defined empirically using optimum thresholds for the scores (T_s) and parameters (T_f) centered around M_s and M_f respectively. These thresholds were obtained as a tradeoff between error and sensitivity (proportion of samples who are correctly identified), since an increase in error improves the sensitivity of the system. The thresholds are obtained empirically. The threshold for scores is two times greater than the parameter threshold. As there are more chances of errors (calibration, recording and sampling errors) in measured variables than scores, so the threshold for scores is larger such that there should be less deviations in scores.

We classified a score as having a deviation if a discrepancy was observed in the score given by a subject (S_i) but not in their physiological parameters F_i for that activity. Summarizing, a deviation in score was marked as true in an activity i if

$$Abs(S_i - M_s) > T_s \qquad (1)$$

$$Abs(F_i - M_f) \leq T_f \qquad (2)$$

The scores for all three surveys were compared to the first principal component of the four features of the EDA, which was used to represent the physiological parameters. In Figures 3, 4 and 5, the self-reporting scores are plotted against the normalized parameter signature of the first principal component of EDA for the memory search task (activity number 2). Figure 3 shows one deviation in the scores as there is one score value which falls outside the score threshold range whereas its normalized EDA component is inside the feature threshold range. In Figure 4, there are two subjective deviations for the difficulty survey whereas for the post-stress survey in Figure 5, there are three subjective deviations.

Table I also presents deviation values for all the three surveys. The deviation values in stress and difficulty scores are 6.9% and 7.2%, respectively whereas post-stress survey contains 10.5% deviation. The percentage in each survey is a summation of the total number of activities which are scored incorrectly. The number of incorrectly scored activities for each subject is determined and a cumulative percentage is obtained across all the subjects. Stress and difficulty in the tasks are closely related to each other hence most of the subjects marked similar scores on these two surveys.

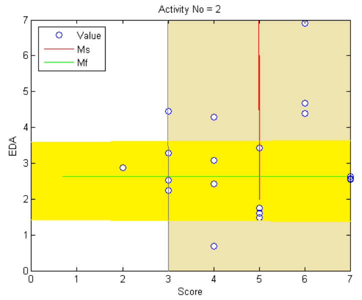

Figure 3. Self-reporting scores and EDA response for the Stress survey. The beige Shaded portion shows the threshold range for scores whereas the yellow shaded portion shows the threshold range for EDA response. Ms and Mf are the median score and median EDA response, respectively.

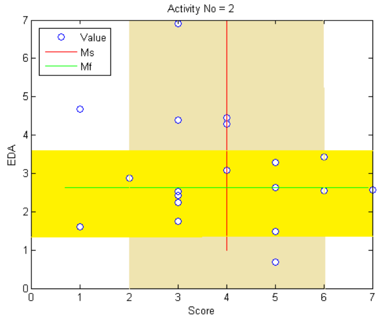

Figure 4. Self-reporting scores and EDA response for Difficulty survey

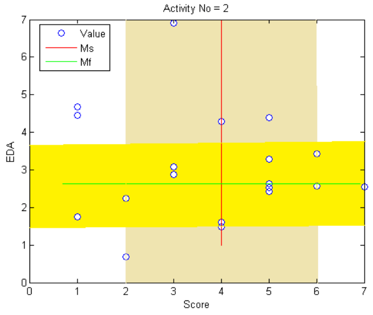

Figure 5. Self-reporting scores and EDA response for Post stress survey

IV. CLASSIFICATION

To validate the correspondence between the perceived stress levels and physiological measures, we developed a classification model using a support vector machine (SVM) based on the radial basis function (RBF) kernel. We used the stress survey to classify the 13 different activities into three classes. The first class contains scores of ones, the second class consists of scores 2, 3 and 4 whereas the third class is based on scores 5, 6 and 7. The assumption was that there should not be any skew in the class samples and classes should contain equal number of samples. The model obtained a classification accuracy of 73%, with the model sensitivity and specificity 72% and 80% respectively.

V. Discussion

The results of self-reporting surveys can be biased due to the varying subjective assessments performed by the subjects. The reasons include low self-confidence, self-biasing and memory recall. The consistency of the conducted surveys was determined using variance across the subjects for each activity and it was found that the difficulty survey has the least variance. Our results show that the EDA signature correlated more strongly with the stress survey than with the other two surveys which suggest that subjects found it easier to score activities in terms of stress rather than in terms of difficulty levels whereas in post-stress survey, recalling the exact stress levels might be one of the reasons for its low performance. An assessment of the accuracy of the self-reporting scores found that the stress survey had the most accuracy. These results indicate that the stress survey is the most effective in accurately determining an individual's appraisal for the stress.

VI. Conclusions

In this paper we have presented an objective comparison of the performance of three different user surveys. We proposed a three-pronged strategy to evaluate the surveys, consistency, relevancy and accuracy. The consistency of a survey was measured by the variance in scores among subjects for a particular activity. The relevancy is based on correlation of a survey to the corresponding EDA response. Finally we measure accuracy to determine the validity of the surveys. With regards to questions in Section I, we found that the stress survey was the most accurate, whereas the post-stress survey had the biasing due to deviations. The scores of the stress surveys were also the most consistent and displayed the highest correlation with physiological variables.

Our results show that stress survey is more accurate and relevant than difficulty and post-stress surveys. The robustness of the system can be improved by developing a model which can minimize individual subjective differences. Future directions include validation of the surveys over longer durations and accurate prediction of self-reporting scores using physiological measures.

Acknowledgment

The work was supported by NPRP grant # [08-125-2-03] from the Qatar National Research Fund which is a member of Qatar Foundation. The statements made herein are solely the responsibility of the authors.

VII. Bibliography

[1] H Seyle, *History and Present status of the stress concept*. NewYork: Free Press: Hnadbook of stress: theoretical and clinical aspects, 1982.

[2] S Kasl, "Stress and Health," *Annual Review of Public Health*, vol. 5, pp. 319-341, 1984.

[3] J Blumenthal, P Stein, L Watkin, L Berkman, S Czakowski, C Oconor, P stone, K Freedland and R Carney, "Depression, heart rate variability and acute myocardial infraction," *Circulation*, vol. 104, no. 17, pp. 2024-2028, Sep 1993.

[4] B McEwen, "Stress, adaptation and disease, allotosis and allostatic load," *Annals of the Newyork Academy of Sciences*, vol. 840, pp. 33-44, 1998.

[5] N Bouger, A Delonges, R Kesler and E Wethington, "The contagion of stress across multiple roles," *Journal of Marriage and Family*, vol. 51, no. 1, pp. 175-184, 1989.

[6] H Mccubbin, C Joy, A Cauble, J Comeau, J Patterson and R Needle, "Family stress and coping: A Decade Review," *Journal of Marriage and Family*, vol. 42, no. 4, pp. 855-864, 1980.

[7] E Jovanov, A ODonnel, D Raskovic, P Cox, R Adhami and F Andresik, "Stress monitoring using a distributed wireless intelligent sensor system," *IEEE Engineering and Biology Magzine*, vol. 22, no. 3, pp. 49-55, June 2003.

[8] J Choi and R Gutierrez-Osuna, "Using heart rate monitors to detect mental stress," in *6th International workshop on wearable and implantable body sensor networks*, Berley CA, June 2009, pp. 219-223.

[9] N Hjortskov, D Rissen, A Blangsted, N Fallentin, U Lundberg, K Sogaard, "The effect of mental stress on heart rate variability and blood pressure during computer work," *European Journal of Applied Physiology*, vol. 92, no. 1, pp. 84-89, 2004.

[10] C Stez, B Anrich, J Schumm, R Marca, G Troster and U Elhlert, "Discriminating stress from cognitive load using a wearable EDA," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 410-417, March 2010.

[11] Bernardi, J Szuk, C Valenti, S Costoldi, C Passino and G Spadacini, "Effects of controlled breathing, mental activity and mental stress with or without verbalization on heart rate variability," *JAmC Cardiology*, pp. 1462-1469, May 2000.

[12] S Cohen, T Kamerck and R Mermelstein, "A Global measure of perceived Stress," *Jounal of Health and Social Behavior*, vol. 24, pp. 38-396, December 1983.

[13] A Adams, S Soumerai, J Lomas and D Degnan, "Evidence of self-report bias in assessing adherence to guidelines," *International Journal for Quality in Health Care*, pp. 187-192, 1999.

[14] L Crockett, J Schulenberg and A Petersen, "Congruence between objective and self-report data in a sample of young adolescents," *Journal of Adolescent Research*, vol. 2, pp. 383-392.

[15] J Choi and R Guiterrez-Osuna, "Removal of respiratory influences from heart rate variability in stress monitoring," *IEEE Sensors Journal*, vol. 11, pp. 2649-2656, 2011.

[16] M Dawson, *The Electrodermal System in Handbook of Psychophysiology.*: Cambridge Univesity Press, 2007.