Chapter 3
Predicting Task Difficulty Through Psychophysiology

Junoš Lukan and Gregor Geršak

Introduction

Psychophysiology investigates changes in the activity of physiology caused by psychological input (Ravaja, 2004). In principle, psychophysiological measuring devices can be divided into two types, the brain scanning apparatuses for central nervous system and devices for autonomous nervous system dynamics' observation. The latter enable measurement of different physiological parameters, e.g. heart rate, heart rate variability, blood pressure, skin conductance, skin temperature, facial thermal scan, breathing rate and breathing amplitude, pupil dilatation, etc. They have been used with increasing regularity to study different constructs, like mental load or effort, stress, emotions, level of focus, difficulty of a task, etc. (Benedek & Kaernbach, 2010a; Collet et al., 1997, 2009; Fauvel et al., 2000; Kivikangas et al., 2014; Olsson & Phelps, 2007; Storm et al., 2005; Wen et al., 2014)

The relationship between physiological measures and task difficulty has been widely studied. The level of arousal can be linked with the level of challenge, focus, and excitement associated with the difficulty of the task (Cacioppo et al., 2007; Lewis et al., 1993; Mandryk & Atkins, 2007). The reports are sometimes contradictory, but some general conclusions can be drawn. In the majority of related work increase in skin conductance (usually called electrodermal activity, EDA) was found with increased arousal (Boucsein, 2012; Boucsein et al., 2012; Brouwer et al., 2013, 2014; Lisetti & Nasoz, 2004). A number of studies reported correlation between EDA and task engagement. In general, EDA increases with difficulty of the task (Clark et al., 2018; Mandryk & Atkins, 2007; Mehler et al., 2009; Nourbakhsh et al.,

J. Lukan (✉)
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia

G. Geršak
Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia

I. Devetak and S. A. Glažar (eds.), *Applying Bio-Measurements Methodologies in Science Education Research*,
https://doi.org/10.1007/978-3-030-71535-9_3

2012; Novak et al., 2014; Pecchinenda, 1996). Heart rate and heart rate variability are subject to effort and workload level associated with task difficulty (Aasman et al., 1987; Veltman & Gaillard, 1998). Similarly, respiration rate increases with mental effort (Butler et al., 2006; Karavidas et al., 2010; Mehler et al., 2009; Veltman & Gaillard, 1998). Skin temperature decreases with arousal due to psychologically induced vasodilatation of peripheral veins (Cacioppo et al., 2007).

In this paper, an attempt of physiology-based prediction of the task difficulty as perceived by the subject is described. Elementary school children were instructed to solve science problems, while their physiology was monitored. Science problems were composed of questions from physics and biology on the state of the matter, microscopic and macroscopic representation of the matter (Slapničar et al., 2017). After each task, they rated the difficulty of the problem. The relationship between its perceived difficulty and physiological parameters was studied.

Methods

Participants

The non-random sample of this pilot study included 10 participants: five were 12 years old (three girls and two boys) and five were 14 years old (three girls and two boys). To ensure anonymity, pupils were assigned a code consisting of a serial number and their age. The subjects were selected from a mixed urban population. They were first familiarised with the purpose and the content of this study and then asked for consent to participate.

Task

The subjects were presented four computer-displayed tasks from the field of science, such as identification of the solid, liquid and gaseous state of water (iceberg, lake, kettle steam), describing the melting of a glacier and opening the gassed beverage bottle, identification of warming of air in a hand pump, estimating water concentration in plants, describing sugar dissolution in water, etc. (Fig. 3.1).

To ensure the least movement possible, subjects were instructed to answer vocally. They could decide from three given choices (Fig. 3.1) and were asked additional questions in case of incomplete or incomprehensible answers. The tasks were designed by three higher education teachers of chemistry and physics and evaluated by six elementary and secondary school teachers (Slapničar et al., 2017). The tasks contained a text presentation (socio-scientific context), visualisations (pictures, schemes, animations) and three-choice answers to be answered by the subjects. After each task, the subjects were asked to grade the properties of the task by a questionnaire of the

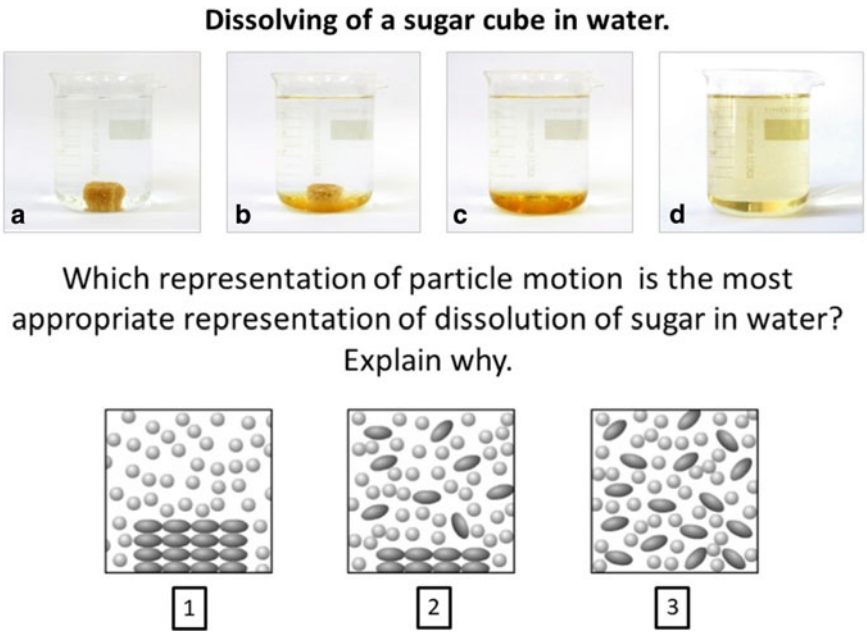

Fig. 3.1 An example of a task from chemistry; dissolving of a sugar cube in water, represented in the lower part using video clips showing movement of the particles

paper-pencil type (five-point Likert-type-scale). The following items were evaluated (1) the difficulty of each task, (2) the confidence of writing the correct answer and (3) whether the task was interesting.

Measuring Protocol

After the initial introduction of the measurement instruments to the test subject, sensors were attached and a rest period was allowed for the initial instructions. This period was set to 5 min to enable the electrolyte gel absorption resulting in an optimal electrical contact (see Ogorevc et al., 2013 for the importance of gel in skin-conductance measurements). Rest period served as a baseline for all the physiological parameters.

The subjects were instructed to sit in a relaxed manner and perform only slow movements if at all. The disturbances from the surroundings (noise from the corridor, activities in the neighbouring rooms) were minimised and were considered negligible. Room temperature and air humidity, the major environmental error sources for skin-conductance measurements, were monitored throughout the experiment. The room's lightning was kept at the same level for all subjects to avoid errors in heart rate estimation by means of photoplethysmography.

This study was approved by the Ethics Committee of the Faculty of Education, University of Ljubljana, resulting in consents obtained for the subjects from school boards, teachers and parents. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Instrumentation

The physiology was acquired by means of a multi-parameter psychophysiology measuring system (MP150 by Biopac, USA) enabling measurements of electrodermal activity, skin temperature, heart rate and respiration with a sampling frequency of 1 kHz. The measuring system was validated in a prior study (Ogorevc et al., 2013). Measuring sites were selected according to Ogorevc et al. (2011) and van Dooren et al. (2012).

Electrodermal activity was monitored by using reusable wet Ag-AgCl electrodes attached to the skin by means of Velcro bands and a Biopac EDA100C amplifier. Measuring sites were the distal phalanges of pointing and middle finger. Skin temperature (ST) was recorded on the ring finger using SKT100C amplifier and a small-size fast response thermistor. Heart rate was calculated from the raw photoplethysmograph signal recorded by a transducer placed on the little finger and connected to amplifier Biopac PPG100C (Fig. 3.2). Respiratory rate was calculated from the

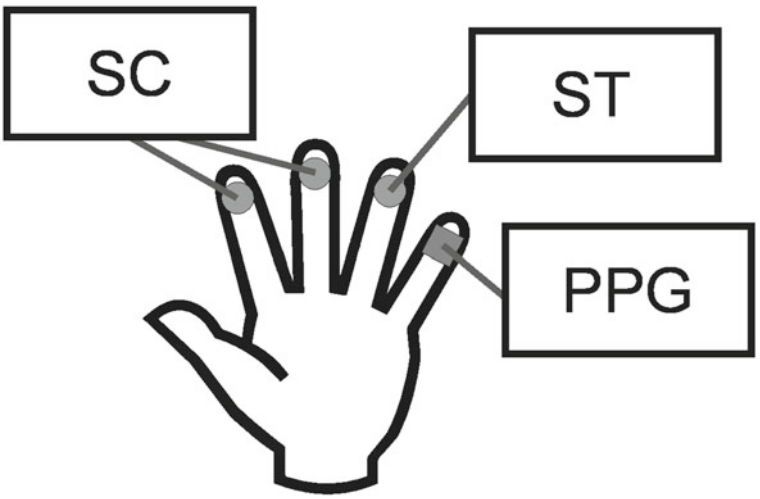

Fig. 3.2 Placement of the physiological sensors on subject's non-dominant hand; SC—skin conductance, ST—skin temperatures, PPG—photoplethysmography for heart rate measurements

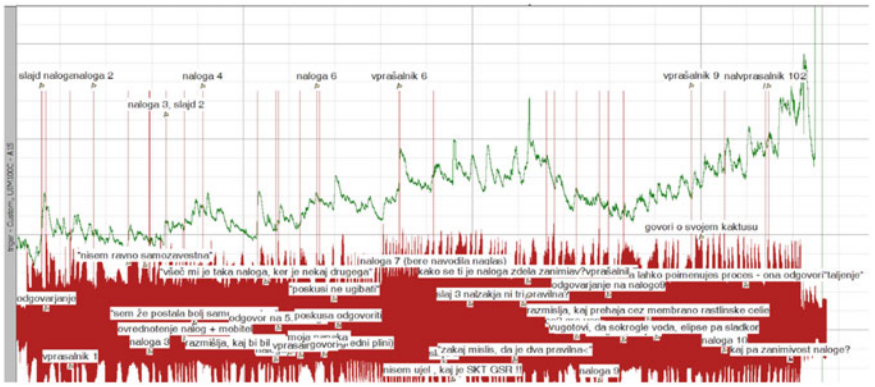

Fig. 3.3 Raw signal of the skin conductance (upper, green curve) with trigger points (red vertical lines) for marking the time span of the tasks. Manual notations are visible in the lower part of the figure, overlaid over the respiration rate signal

acquired signal of the subject's chest displacement while breathing by means of a chest-belt transducer connected to Biopac RSP100C amplifier.

In addition, an operator used a manual trigger for timestamping the beginning and end of each task and of answering the questionnaires. Manual notations were possible for additional information like unexpected events, observation of subject's behaviour and emotional state, plausible explanation of sudden physiological change, etc. These are visible in the lower part of Fig. 3.3.

Psychophysiological parameters were monitored throughout the task and the subsequent filling out of the questionnaire. The physiological signals were recorded using the software AcqKnowledge 4.1 (Biopac, 2014). The acquired raw signals were pre-processed (filtered, outliers removed) and stored for further signal processing.

Signal Processing

The photoplethysmogram was pre-processed using the AcqKnowledge software. First, a bandpass filter was applied with frequency cut-offs at 0.5 Hz and 3 Hz, using a Hamming window. Next, the AcqKnowledge's in-built cycle detector was employed, where cycles were determined from peaks and a specific threshold configuration was determined for every subject. The peaks' locations were then exported to a `CSV` file.

Heart rate and heart rate variability were analysed using an `R` package `RHRV` (Rodriguez-Linares et al., 2016). Effectively, the raw heart rate signal was first extracted from heartbeat positions previously exported from AcqKnowledge. Next, this signal was further filtered to only include heart rate between 50 bpm and 140 bpm. The signal was then interpolated using spline interpolation. Finally, a Fourier transform was calculated by shifting a 300-second window 0.5 s at a time. The large

Table 3.1 Frequency limits as specified in heart rate variability calculation. These roughly correspond to the limits suggested by European Society of Cardiology (Malik, 1996; see page 360, Table 2) with a wider ULF band

Frequency band	Lower frequency limit (Hz)	Upper frequency limit (Hz)
Ultra low frequency (ULF)	0.00	0.03
Very low frequency (VLF)	0.03	0.05
Low frequency (LF)	0.05	0.15
High frequency (HF)	0.15	0.4

window chosen here places a lower limit of 48 bpm on heart rate due to the Nyquist–Shannon sampling theorem (Clifford, 2002). Thus, a time dependence of heart rate was obtained with a resolution of 0.5 s and heart rate variability was calculated in bands as specified in Table 3.1.

Respiration rate was analysed similarly to heart rate. Since the package RHRV is intended for heart rate analysis and some parameters are set in the source code, the breath time positions were first divided by 3. Using this linear transformation, the respiration rate signal is within pre-specified limits of the heart rate analysis. The time dependence of the respiration rate was finally divided by 3 again to account for the previous transformation. Then the signal's mean, standard deviation and maximum and minimum values were calculated.

Mean, standard deviation, maximum and minimum values, mean slope and the last value of skin temperature were also calculated. The last value was added because skin temperature changes are relatively slow due to skin's slow thermal response to psychological stressors and due to high tissue heat-capacity. The latest temperature value during the task was expected to be more representative of the psychological effect the task had on the subject (Novak et al., 2011).

Skin conductance was analysed using a `MATLAB`-based program `Ledalab` (Benedek, 2014). Specifically, its continuous decomposition analysis and the traditional trough-to-peak analysis were chosen (Benedek & Kaernbach, 2010a cf. Benedek & Kaernbach 2010b for (nonnegative) discrete decomposition). The software is intended for analysing a specific time interval after an event. To accommodate our experimental design, one of their scripts, called `export_era.m`, was modified so that the pre-set time intervals corresponding to experimental conditions could be analysed. Table 3.2 lists the skin-conductance feature calculated using `Ledalab`. The first one is a classical trough-to-peak method (TTP), which simply counts the peaks as defined by a minimum amplitude criterion and determines their amplitudes by measuring the distance from the peak to the preceding trough. A more complex method, continuous decomposition analysis (CDA), takes a deconvolution of the signal and thus decomposes it into tonic activity (i.e. skin-conductance level, SCL) and phasic activity (which corresponds well to the skin-conductance responses, SCR). Finally, since the number of SCRs and the time integral of the phasic driver are all dependent on the task duration, these three features were divided by the duration of the task.

Table 3.2 Skin-conductance features and their meanings as calculated using Ledalab (Benedek, 2014; Benedek & Kaernbach, 2010b). The prefix of the feature denotes the method used: continuous decomposition analysis (CDA) and the standard trough-to-peak (TTP) method

Feature	Meaning
CDA.nSCR	Number of skin-conductance responses (SCRs) within the task
CDA.Latency	Time to the first SCR within the task
CDA.AmpSum	Sum of amplitudes of SCRs within the task
CDA.SCR	Mean phasic driver, which corresponds to the average of SCRs' amplitudes
CDA.ISCR	Time integral of the phasic driver
CDA.PhasicMax	Maximum of phasic activity, corresponding to the max. SCR amplitude
CDA.Tonic	Mean tonic activity, corresponding to mean SCL
TTP.nSCR	Number of SCRs within the task
TTP.AmpSum	Sum of amplitudes of SCRs within the task
TTP.Latency	Time to the first SCR within the task
Global.Mean	Mean of the skin-conductance (SC) signal
Global.MaxDeflection	Maximum positive deflection in the SC signal

Results

Self-Reports on Perceived Difficulty

Before attempting a psychophysiological analysis, the subjects' self-reports were analysed. Bivariate relationships between the psychological variables were observed in scatter plots and the strongest relationship was found between the subjects' perceived difficulty of the task and the perceived accuracy of their answer. The harder the task was perceived as, the less certain they were that they solved it correctly, the Pearson correlation coefficient was $r = -0.58$ with a $p < 0.001$. The remaining relationships were less apparent with some being non-linear. Such was the relationship between the task order and its perceived difficulty, where later tasks were assessed as more difficult (the Spearman correlation coefficient had a value of $\rho = 0.45$, $p < 0.001$), but not consistently. Other correlations had $|\rho| < 0.4$, albeit some were statistically significant. It is also worth noting that the correlation between the comprehension of the task and its perceived difficulty was low ($\rho = -0.20$, $p = 0.040$) and that the rating of the comprehension was high on average (mean comprehension across all subjects and all tasks was out 4.65 of 5).

Thus it was determined that perceived difficulty of the task was a sufficiently distinct psychological construct, not identical to the solver's comprehension, engagement with the task or conviction about their own answer. The task difficulty was therefore sought to be predicted by physiological parameters.

Choosing Parameter's Best Features

In attempting to answer how to best predict task perceived difficulty based on physiological reactions, a measure of each physiological process was first needed to be chosen. Different physiological parameters (such as skin conductance) were analysed according to different measures (such as the number of skin-conductance responses and the mean skin-conductance level). These were sometimes unrelated or relatively independent features, but could also be the same physical attributes, but calculated according to different methods (cf. skin-conductance measures in Table 3.2). Therefore, a decision was first made as to the best way to assess each of the measured parameters, since feature selection was found to be important in other studies (e.g. Kukolja et al., 2014).

To determine the best predictor, linear regression was employed. The predictors were included in the model step-wise and the correlations between them, their β coefficients and the Akaike information criterion (AIC) were taken into account when deciding which ones to include and which ones to keep in the model. Only simple linear effects and no interactions were considered in regression models in this part of the analysis.

Both, the physiological and psychological variables were linearly transformed before the following analysis. Specifically, they were standardised within each subject, so that the mean of every variable within the subject was 0 and its standard deviation was 1. This was done in order to circumvent the problem of different baselines and to make the distributions closer to the normal distribution.

Skin Temperature

As noted, mean, standard deviation, maximum and minimum values, mean slope and the last value of the temperature were calculated. The predictor that explained the most variance of the difficulty was the mean skin temperature. Compared to the zeroth-order model it lowered the Akaike information criterion (AIC) by the largest amount and was also the most statistically significant predictor in the full model, containing all of the calculated features. The model with only the mean temperature as the predictor was statistically significant ($F(1, 104) = 10.3$, $p = 0.002$) with the beta regression coefficient $\beta = 0.300$ ($t = 3.21$, $p = 0.002$). The mean skin temperature explained 8.1% of the difficulty variance, compared to the adjusted $R^2_{\text{adj}} = 0.19$ of the full model. Its correlation was the highest with the minimum ($r = 0.56$) and the final value of skin temperature ($r = 0.55$, both $p < 0.001$). On the other hand, the information about the standard deviation and the slope might be lost with this model since the correlations between them and the chosen feature are $r < 0.1$ and indistinguishable from zero.

Respiration Rate

The respiration rate feature that lowered the AIC the most was the maximum value of the respiration rate during the task. If all of the calculated features were included as predictors in linear regression, the standard deviation of the respiration rate had the highest absolute value of the β-coefficient. This suggests that these two predictors are correlated, which was indeed the case ($r = 0.47$, $p < 0.001$). It should be noted, however, that the predictions took opposite directions. Specifically, a greater maximum value of respiration rate was related to higher task difficulty, while there was more deviation in respiration rate at lower difficulties.

Of these two, the maximum value seems to be the simpler feature. It is also, however, more prone to moving artefacts and thus less reliable. The standard deviation is the simplest indicator of the respiration rate variability. Since the correlations of these two features with task difficulty were in opposite directions, it was decided to keep both as predictors in the subsequent analysis. Together, they explained over 11% of difficulty variance.

Heart Rate

Simple heart rate measures and heart rate variability as calculated using Fourier transform were separated for the purpose of statistical analysis. The simple features of heart rate showed a very similar pattern to the features of the respiration rate. The two predictors that lowered the AIC the most were the maximum value of the heart rate and its (simple) standard deviation. They were also found to be correlated ($r = 0.52$, $p < 0.001$) and their relation to the task difficulty was in opposite directions. The maximum value, however, was the strongest predictor in the full model ($\beta = 0.487$, $t = 4.72$, $p < 0.001$ for maximum value compared to the $\beta = -0.402$, $t = -3.57$, $p < 0.001$ for the standard deviation).

The power in different regions of the power spectrum (see Table 3.1) showed moderately strong linear relationships. It was tested whether the heart rate variability calculated in this fashion could predict task difficulty and replace the simple standard deviation. However, neither the full model containing power in all of the regions in the power spectrum nor the model with only the strongest (by the information criterion) predictor, the power in the very low frequency band (VLF), were statistically significant.

It was thus decided to keep both the heart rate maximum and the simple standard deviation (as a simple measure of heart rate variability) as the predictors for further analysis, since they were only moderately correlated and they explained over 16% of the difficulty variance.

Skin Conductance

Many skin-conductance features were calculated by different methods (see Table 3.2). The predictors were first separated into three blocks, according to the method used. They were then analysed separately and only the best predictors from each method were considered for further analysis. Of the calculated features, both global features (mean and maximum deflection) turned out to be good parameters. The number of SCRs and the sum of their amplitudes as calculated by trough-to-peak method (TTP.nSCR and TTP.AmpSum) were also statistically significant. Among those calculated by the continuous decomposition analysis, four of the predictors were statistically significant: the sum of SCR amplitudes (CDA.nSCR), the time integral of the phasic driver (CDA.ISCR) and its maximum amplitude (CDA.PhasicMax) and the mean tonic activity (CDA.Tonic).

Some of these predictors represent the same physiological processes and this fact was reflected in high correlations. Indeed, the correlations between the number of SCRs as calculated by two different methods, between the sums of SCR amplitudes, and between the mean tonic component and a simple average of the skin-conductance signal were all nearly perfect. It was thus prudent to keep only one of each pair in the final model. The choice was made by determining which of the two lowered the AIC more. It was found that in these three pairs, the features calculated by the continuous decomposition analysis were superior in their predictive power compared to the ones calculated by other methods.

Finally, the best features among those calculated by CDA were chosen. Three predictors were found to be statistically significant: the sum of amplitudes (CDA.AmpSum), the maximum amplitude (CDA.PhasicMax) and the time integral of the SCRs (CDA.ISCR). Naturally, the maximum SCR amplitude and the sum of all of them were correlated ($r = 0.62$, $p < 0.001$), which resulted in a higher variance inflation factor ($VIF \approx 1.7$) for both. Of these two features only the sum of amplitudes was therefore chosen for further analysis, since it is less susceptible to moving artefacts compared to a point feature and also had a higher β-coefficient. The sum of amplitudes and the time integral of the SCRs together explained more than 33% of variability in task difficulty.

Choosing the Best Parameter

In the previous section, features of individual parameters were explored. Specifically, their predictive power pertaining to task difficulty was analysed. After one or two features of each parameter were chosen, the aim was to construct a regression model consisting of any number of physiological parameters.

The predictors included in the full physiological model were:

- mean skin temperature,
- the maximum value and standard deviation of respiration rate,

- standard deviation of heart rate and the maximum heart rate and
- sum of skin-conductance responses amplitudes and their time integral.

Most of these predictors were correlated with the task difficulty, but only moderately correlated between themselves: the highest correlation remained that between the maximum heart rate and its standard deviation ($r = 0.52$, $p < 0.001$). Several possibilities for a regression model were therefore tested, all constructed in a 'backwards' way, eliminating predictors one by one.

In the full regression model, consisting of all physiological predictors listed above, all but the maximum respiration rate ($\beta = 0.106$, $t = 1.16$, $p = 0.250$) were statistically significant. Furthermore, eliminating any predictor (other than the maximum respiration rate mentioned) only increased the Akaike information criterion. The predictors with the highest β-coefficients were both related to skin-conductance responses, with the sum of SCRs amplitudes having $\beta = 0.444$ ($t = 4.45$, $p < 0.001$) and the time integral of the SCRs $\beta = -0.388$ ($t = -4.55$, $p < 0.001$). A model that included all chosen predictors had an $R^2_{\text{adj}} = 0.450$ and so it explained almost half of the task difficulty variability.

A model that included only the best two predictors, namely the amplitude and the time integral of the SCRs, fared significantly worse than the full model ($F(97, 92) = 4.17$, $p = 0.002$). However, it still explained over a third of the variance of task difficulty ($R^2_{\text{adj}} = 0.361$).

Effect of Task Duration

An attempt was made to explore the role of task duration. Including task duration as a predictor in the full physiological model rendered all other predictors insignificant. Indeed, predicting task difficulty using task duration as a sole predictor in a linear regression model explained more than a half of the task difficulty variance ($R^2_{\text{adj}} = 0.506$). Due to collinearity with duration, the variance of other regression coefficients is increased: the most so for the predictors related to SCRs. The *VIF*s for the sum of SCRs amplitudes, their time integral and task duration were 3.49, 2.14 and 3.26, respectively.

To diminish the effect of task duration in the model, another, limited, physiological model was considered. It excluded the features of SCRs and only included mean skin temperature, standard deviation of respiration rate and its maximum and standard deviation of heart rate and its maximum. In this model, all predictors were statistically significant and the model had $R^2_{\text{adj}} = 0.296$.

The collinearity of the task duration and the SCR features might be able to illuminate the nature of skin-conductance response, however. The time integral of the SCRs and the sum of their amplitudes are, theoretically, related to the task duration. The first relationship should be monotonously increasing: the longer the task, the more SCRs are generally expected. This means that the sum of their amplitudes, too, is increasing with time. The second relationship is less straightforward. The time integral is broadly expected to increase with task duration. In contrast with discrete

(nonnegative) deconvolution method (Benedek & Kaernbach, 2010b), however, the phasic driver representing the SCRs can be negative as well as positive when using continuous decomposition analysis. This means that the time integral of the phasic driver does not increase with time in a simple monotonous manner.

These two different relationships between two features of skin conductivity and task duration are reflected in different correlation coefficients we measured. While the sum of SCRs amplitudes correlated moderately with task duration ($r = 0.45$, $p < 0.001$), the correlation between task duration and the time integral of the phasic driver was not statistically significant ($r = -0.18$, $p = 0.073$).

Despite this, the nature of skin-conductance response cannot be unambiguously inferred from our data. The sum of SCRs amplitudes was higher during the tasks of higher perceived difficulty ($\beta = 0.648$, $t = 7.34$, $p < 0.001$, in a model consisting of only this predictor and the CDA.ISCR). It remains unclear, however, whether this is due to a larger number of SCRs (the time-normalised number of SCRs was not a good predictor in a model consisting of skin conductivity features, $\beta = -0.115$, $t = -1.41$, $p = 0.257$) or to their higher average amplitude. Additionally, the evidence was inconclusive regarding the relationship between the shape of the SCRs and the task difficulty, since the integral of the phasic driver had a negative β coefficient but was not statistically significantly correlated with difficulty when the effect of the sum of amplitudes was not controlled for.

Discussion

In the previous chapter, features of several physiological parameters were considered as predictors of task difficulty. This was done both, by considering the β coefficients of the individual predictors in a comprehensive model and by testing smaller models consisting of only selected features.

Regardless of the physiological parameter under consideration, it was possible to build a statistically significant regression model. Different parameters had different predictive power for task difficulty, however.

In terms of the proportion of explained variance, skin conductivity and its features would seem to be the best physiological parameter for predicting task difficulty, at least judging naively by the R^2_{adj} coefficient of a model composed of two of its features. Taking into account their relationship with the task duration, this conclusion is less convincing.

Task duration was the strongest predictor of task difficulty when comparing it to selected physiological features. It did not render the physiological features redundant, however. Even the features related to skin-conductance responses might still hold valuable information, since their time dependence remains unclear (see "Effect of Task Duration").

Other physiological parameters were good predictors of task difficulty even when skin conductance was not included in the regression model; they explained almost 30% of the variance of the task difficulty. Of these, the best predictors were the

maximum heart rate and the mean skin temperature during the task. Indeed, by themselves they explained 11 and 8% of the variance, respectively, when considering regression models consisting of only one predictor.

It would seem then that an attempt to predict how difficult a task is from physiological processes could take several different forms. One way would be to only measure skin conductivity and infer task difficulty from SCRs: specifically, from their amplitude sum and their time integral. One should consider, however, what are the benefits of this set-up compared to simply timing the task and concluding about its difficulty from its duration. More specifically, the relationship between the SCRs features and task duration should be considered before determining whether this would be a fruitful approach.

An alternative approach would be to measure heart and respiration rate and skin temperature, instead, and using appropriate features to predict difficulty. It could be argued that this provides additional information about the task difficulty compared to a simple duration measurement. Furthermore, task duration is a *post festum* measure, while physiological features could in theory be calculated concurrently with solving of the problem of which difficulty is predicted.

There are other factors to consider when choosing a physiological process for task difficulty prediction. The two physiological parameters from which the best predictors were chosen—skin conductivity and heart rate, measured by photoplethysmography—are, arguably, also the most delicate. First, they are both sensitive to moving artefacts and demand careful attachment of the transducers. In addition, skin conductance suffers from errors due to non-stable electrical contact between electrodes and the skin. Secondly, to extract individual heartbeats and skin-conductance responses, significant (pre)processing of the signals is required. This often demands some manual inspection of the signal or results in erroneous detections. Furthermore, doing such analysis on the fly would require considerably more processing power. Finally, while the heart rate measurements are readily accessible via wearable instruments such as smart watches and bracelets, skin conductivity on the other hand is less commonly reported.

There are several limitations to the present study. The small size of the sample made it difficult to assess the distribution of individual physiological features. Assumptions of normality could therefore not be reliably tested. This was partly compensated for by transforming (standardising) the data. In addition, the design of the experiment was such that the sequence of the tasks was not independent of their difficulty. There was a moderate correlation between the task order number and its difficulty (Spearman's $\rho = 0.45$, $p < 0.001$), but the relationship was non-linear and was not accounted for in regression analysis.

In conclusion, the results of our study showed that the perceived difficulty of a task could be predicted by measuring physiological processes and calculating some of their features. Future work could be focused on determining more reliably what the most advantageous processes and features are and to establish the nature of this relationship more definitely.

Acknowledgements The authors acknowledge that the study was part of the project, Explaining Effective and Efficient Problem Solving of the Triplet Relationship in Science Concepts Representations (J5-6814), which was financially supported by the Slovenian Research Agency.

The authors declare that they have no conflict of interest.

References

Aasman, J., Mulder, G., & Mulder, L. J. M. (1987). Operator effort and the measurement of heart-rate variability. *Human Factors, 29*(2), 161–170. https://doi.org/10.1177/001872088702900204.

Benedek, M. (2014). *Ledalab*. Institut für Psychologie, University of Graz, Austria; University of Graz. Retrieved from http://www.ledalab.de/.

Benedek, M., & Kaernbach, C. (2010a). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods, 190*(1), 80–91. https://doi.org/10.1016/j.jneumeth.2010.04.028.

Benedek, M., & Kaernbach, C. (2010b). Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology, 47*(4), 647–658. https://doi.org/10.1111/j.1469-8986.2009.00972.x.

Biopac. (2014). *MP system hardware guide* (Ver. 21. 1). Goleta, CA: BIOPAC Systems. Retrieved August 10, 2015, from http://www.biopac.com/Manuals/mp_hardware_guide.pdf.

Boucsein, W. (2012). *Electrodermal activity* (2nd ed., p. 618). New York City: Springer Science and Business Media.

Boucsein, W., Fowles, D. C., Grimnes, S., Ben-Shakhar, G., Roth, W. T., Dawson, M. E., & Filion, D. L. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology, 49*, 1017–1034. https://doi.org/10.1111/j.1469-8986.2012.01384.x.

Brouwer, A. M., Hogervorst, M. A., Holewijn, M., & van Erp, J. B. (2014). Evidence for effects of task difficulty but not learning on neurophysiological variables associated with effort. *International Journal of Psychophysiology, 93*(2), 242–252. https://doi.org/10.1016/j.ijpsycho.2014.05.004.

Brouwer, A.-M., van Wouwe, N., Mühl, C., van Erp, J., & Toet, A. (2013). Perceiving blocks of emotional pictures and sounds: Effects on physiological variables. *Frontiers in Human Neuroscience, 7*(June), 1–10. https://doi.org/10.3389/fnhum.2013.00295.

Butler, E. A., Wilhelm, F. H., & Gross, J. J. (2006). Respiratory sinus arrhythmia, emotion, and emotion regulation during social interaction. *Psychophysiology, 43*(6), 612–622. https://doi.org/10.1111/j.1469-8986.2006.00467.x.

Cacioppo, J., Tassinary, L. G., & Berntson, G. G. (Eds.). (2007). *The handbook of psychophysiology* (3rd ed., p. 914). New York: Cambridge University Press. https://doi.org/10.1017/cbo9780511546396.

Clark, D. J., Chatterjee, S. A., McGuirk, T. E., Porges, E. C., Fox, E. J., & Balasubramanian, C. K. (2018). Sympathetic nervous system activity measured by skin conductance quantifies the challenge of walking adaptability tasks after stroke. *Gait and Posture*, *60*(August 2017), 148–153. https://doi.org/10.1016/j.gaitpost.2017.11.025.

Clifford, G. D. (2002). *Signal processing methods for heart rate variability analysis.* Doctoral dissertation. University of Oxford. Retrieved from http://www.ibme.ox.ac.uk/research/biomedical-signal-processing-instrumentation/prof-l-tarassenko/publications/pdf/gdcliffordthesis.pdf.

Collet, C., Averty, P., & Dittmar, A. (2009). Autonomic nervous system and subjective ratings of strain in air-traffic control. *Applied Ergonomics, 40*(1), 23–32. https://doi.org/10.1016/j.apergo.2008.01.019.

Collet, C., Vernet-Maury, E., Delhomme, G., & Dittmar, A. (1997). Autonomic nervous system response patterns specificity to basic emotions. *Journal of the Autonomic Nervous System*, *62*(1–2), 45–57. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9021649.

Fauvel, J. P., Cerutti, C., Quelin, P., Laville, M., Gustin, M. P., Paultre, C. Z., & Ducher, M. (2000). Mental stress-induced increase in blood pressure is not related to baroreflex sensitivity in middle-aged healthy men. *Hypertension*, *35*(4), 887–91. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10775556.

Karavidas, M. K., Lehrer, P. M., Lu, S.-E., Vaschillo, E., Vaschillo, B., & Cheng, A. (2010). The effects of workload on respiratory variables in simulated flight: A preliminary study. *Biological Psychology, 84*(1), 157–160. https://doi.org/10.1016/J.BIOPSYCHO.2009.12.009.

Kivikangas, J. M., Kätsyri, J., Järvelä, S., & Ravaja, N. (2014). Gender differences in emotional responses to cooperative and competitive game play. *PLoS ONE, 9*(7), https://doi.org/10.1371/journal.pone.0100318.

Kukolja, D., Popović, S., Horvat, M., Kovač, B., & Ćosić, K. (2014). Comparative analysis of emotion estimation methods based on physiological measurements for real-time applications. *International Journal of Human Computer Studies, 72*(10–11), 717–727. https://doi.org/10.1016/j.ijhcs.2014.05.006.

Lewis, M., Haviland-Jones, Barrett, J. M., & Feldman, L. (1993). *Handbook of emotions* (p. 720). New York: The Guilford Press.

Lisetti, C. L., & Nasoz, F. (2004). Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP Journal on Advances in Signal Processing, 2004*(11), 1672–1687. https://doi.org/10.1155/S1110865704406192.

Malik, M., Bigger, J. T., Camm, A. J., Kleiger, R. E., Malliani, A., Moss, A. J., & Schwartz, P. J. (1996). Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *European Heart Journal, 17*(3), 354–381. https://doi.org/10.1093/oxfordjournals.eurheartj.a014868.

Mandryk, R. L., & Atkins, M. S. (2007). A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International Journal of Human-Computer Studies, 65*(4), 329–347. https://doi.org/10.1016/J.IJHCS.2006.11.011.

Mehler, B., Reimer, B., Coughlin, J., & Dusek, J. (2009). Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record: Journal of the Transportation Research Board, 2138,* 6–12. https://doi.org/10.3141/2138-02.

Nourbakhsh, N., Wang, Y., Chen, F., & Calvo, R. A. (2012). Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Australian Computer-Human Interaction Conference* (pp. 420–423).

Novak, D., Beyeler, B., Omlin, X., & Riener, R. (2014). Workload estimation in physical human-robot interaction using physiological measurements. *Interacting with Computers*. https://doi.org/10.1093/iwc/iwu021.

Novak, D., Mihelj, M., & Munih, M. (2011). Psychophysiological responses to different levels of cognitive and physical workload in haptic interaction. *Robotica, 29*(3), 367–374. https://doi.org/10.1017/S0263574710000184.

Ogorevc, J., Geršak, G., Novak, D., & Drnovšek, J. (2013). Metrological evaluation of skin conductance measurements. *Measurement: Journal of the International Measurement Confederation*, *46*(9), 2993–3001. https://doi.org/10.1016/j.measurement.2013.06.024.

Ogorevc, J., Podlesek, A., Geršak, G., & Drnovšek, J. (2011). The effect of mental stress on psychophysiological parameters. *IEEE International Symposium on Medical Measurements and Applications, 2011,* 294–299. https://doi.org/10.1109/MeMeA.2011.5966692.

Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature Neuroscience, 10*(9), 1095–1102. https://doi.org/10.1038/nn1968.

Pecchinenda, A. (1996). The affective significance of skin conductance activity during a difficult problem-solving task. *Cognition and Emotion, 10*(5), 481–504. https://doi.org/10.1080/026999396380123.

Ravaja, N. (2004). Contributions of psychophysiology to media research: Review and recommendations. *Media, 6*(2), 193–235. https://doi.org/10.1207/s1532785xmep0602.

Rodriguez-Linares, L., Vila, X., Lado, M. J., Mendez, A., Otero, A., & Garcia, C. A. (2016). RHRV: Heart rate variability analysis of ECG data. Retrieved from https://cran.r-project.org/package=RHRV.

Slapničar, M., Devetak, I., Glažar, A. S., & Pavlin, J. (2017). Identification of the understanding of the states of matter of water and air among slovenian students aged 12, 14 and 16 years through solving authentic tasks. *Journal of Baltic Science Education, 16*(3), 308–323.

Storm, H., Shafiei, M., Myre, K., & Raeder, J. (2005). Palmar skin conductance compared to a developed stress score and to noxious and awakening stimuli on patients in anaesthesia. *Acta Anaesthesiologica Scandinavica, 49*(6), 798–803. https://doi.org/10.1111/j.1399-6576.2005.00665.x.

van Dooren, M., de Vries, J. J. G. G. J., & Janssen, J. H. (2012). Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. *Physiology & Behavior, 106*(2), 298–304. https://doi.org/10.1016/j.physbeh.2012.01.020.

Veltman, J. A., & Gaillard, A. W. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics, 41*(5), 656–669. https://doi.org/10.1080/001401398186829.

Wen, W., Liu, G., Cheng, N., Wei, J., Shangguan, P., & Huang, W. (2014). Emotion recognition based on multi-variant correlation of physiological signals. *IEEE Transactions on Affective Computing, 5*(2), 126–140.

Junoš Lukan has an M.A. degree in Psychology from the University of Ljubljana and an M.Sc. degree in Physics from the Imperial College London. He is a Ph.D. student at the Jožef Stefan International Postgraduate School and a researcher in the Ambient Intelligence Group at the Department of Intelligent Systems at Jožef Stefan Institute, Ljubljana, Slovenia.

Gregor Geršak, Ph.D. is an Associate Professor lecturing measurements, measuring methods, and instrumentation at the Faculty of Electrical Engineering (FE), University of Ljubljana, Slovenia. Apart from classical metrology, his main research area is psychophysiology used in human–computer interaction, affective computing, education, and cognitive science. He makes use of expertise in metrology and calibration, metrology of biomedical sensors and instrumentation, wearables, signal processing, medical thermography, and virtual reality. He was a visiting scholar at Physikalisch-Technische Bundesanstalt, Braunschweig and University of California, Berkeley.