EMOPAIRCOMPETE - PHYSIOLOGICAL SIGNALS DATASET FOR EMOTION AND FRUSTRATION ASSESSMENT UNDER TEAM AND COMPETITIVE BEHAVIOURS

Sneha Das & Nicklas Leander Lund & Carlos Ramos González & Line Clemmensen
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Kongens Lyngby, Denmark 2800
{sned,lkhc}@dtu.dk

Nicole Nadine Lønfeldt
Child and Adolescent Mental Health Center, Mental Health Services
Capital Region of Denmark, 2100
nicole.nadine.loenfeldt@regionh.dk

ABSTRACT

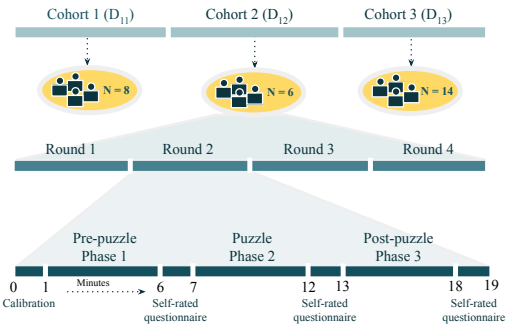
We introduce a new dataset for emotion and stress (frustration) detection from physiological signals. Physiological signals are relevant in health applications like stress detection and monitoring or in mental health where emotion regulation often is of importance. The dataset contributes with the possibilities for disentangling a detailed emotional space (10 emotions) in relation to physiological signals, study dyads of prosocial behaviours vs teams of aggressive behaviours, and investigate continual learning or replication uncertainties [1].

1 INTRODUCTION

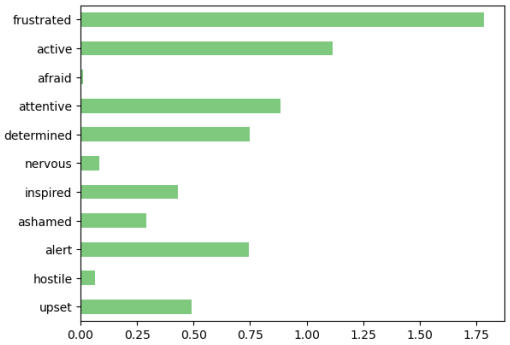
Extensive research has been dedicated to the detection and prediction of stress and relevant health markers from biosignals and wearable devices provide an easy, low-cost, unobtrusive way for real-time monitoring of physiological changes (biosignals) Giannakakis et al. (2019); Schmidt et al. (2019). However, there are lingering challenges in the safe deployment of the technology in healthcare and clinical settings Monteith et al. (2022). Often, the developed models are trained on data collected from highly controlled setting. Controlled studies are often limited and may not generalise as individual responses to stimuli can differ and experimental data is limited in sample size. Other challenges include consistently reproducing model performance under real-world circumstances, within individuals across measurement times and across individuals.

Understanding how emotions and frustration manifest physiologically during teamwork and competition is crucial within mental health treatment Olesen et al. (2023). Most datasets on emotion assessment from physiological signals often focus on individual experiences or lack specific contexts like teamwork and competition Sharma et al. (2019); Park et al. (2020); Bizzego et al. (2021). This limits the development of tools that can understand emotional dynamics in social interactions. This paper addresses the need for a new dataset by presenting EmoPairCompete. This dataset offers physiological signals and emotional annotations through self-rated questionnaires, collected after each phase of the competitive tasks. It includes data from multiple cohorts, offering insights into individual and group-level emotional dynamics. Furthermore, the data is collected in a semi-controlled setting, and hence is a contribution towards the infrastructure for in-the-wild studies and a step towards safer tools for healthcare.

[1] Data is available at GitHub



(a) Illustration of the data cohorts with a timeline of the experiment design over an arbitrary round.



(b) Average difference of responses between the task and rest phases, computed as the difference between task and rest phase. Positive values indicate higher values during task phase than in rest phase.

2 DATA COLLECTION METHODOLOGY

Task: To elicit emotions, we used a game-elicited-emotion paradigm. The tangram task was originally designed to study parent and social behaviour between parents and children Hudson & Rapee (2001). It was designed to be difficult and impossible for the child to finish within the allotted time. A variation of the tangram task has emerged, designed to assess aggressive and prosocial behaviour in adults Saleem et al. (2015). Our task was designed to be difficult to complete in the allotted five minutes and to elicit helpful or prosocial behaviour within the dyad by calling them teams and creating a competition between the teams. The task also necessitated a lot of communication between team members. Participants were divided into pairs of two of their own choosing. Each team was given a 7-piece tangram puzzle and a set of sketches of the puzzle solutions. The teams designated one of the two members as a 'puzzler', and this team member was allowed to touch the puzzle pieces and was responsible for assembling the pieces to solve the puzzle. The other member was the designated 'instructor', who was allowed to see the puzzle solution and was tasked with instructing the puzzler in how to assemble the pieces but was not allowed to touch the puzzle pieces. Also, *only* the final solution to each puzzle, and not the path to the solution, was provided to the instructor. The solved puzzle yielded one point and the team with the most points at the end of round four won. During the competition the participants were asked to rate their emotions in a physical questionnaire. A round consisted of seven parts: calibration, pre-puzzle (phase 1) and puzzle (phase 2), post-puzzle (phase 3), and each phase was followed by a recording of emotions (self-rated questionnaire), as illustrated in Fig. 1a.

Questionnaires: After each condition (rest, stress, recovery), we assessed emotions with the international PANAS Short Form (I-PANAS-SF), which includes five positive emotions (active, alert, attentive, determined, inspired) and five negative emotions (afraid, ashamed, hostile, nervous, upset) that participants rate how much they were experiencing each emotion on a five-point scale from 1 (not at all) to 5 (a lot) Thompson (2007). We also assessed frustration level using a visual analogue scale ranging from 0 (not at all frustrated) to 10 (extremely frustrated) after each condition. Finally, participants rated how difficult they found the task on a visual analogue scale ranging from 0 (not at all difficult) to 10 (extremely difficult) after the stress condition.

Data acquisition: Equipped with the Empatica E4 wearable Emp (2023) during the competition, we measured the electrodermal activity (EDA), heart rate (HR), blood volume pulse (BVP) and temperature of the participants. The participants were instructed to turn on the E4 at the beginning of each round, tag the start and stop times of the rest and puzzle phases within each round, and to turn off the E4 at the end of each round. By doing so, four measurements of $\sim$ 19 min were obtained per participant. Thereby, a total of $\sim$ 72 min of measurement was obtained from each participant. Data were collected over multiple acquisition cohorts (D_{11}, D_{12}, D_{13}) with identical experimental design, conducted at different times over a year. This study was approved by the institutional ethical

review board [Anonymuous REF][2] and participants received written and verbal information about the study before they signed consent forms.

Measuring device: Empatica E4 Wristband

To acquire the data in the experiments, the participants wear an Empatica E4 Wristband. This wearable device contains the following sensors Emp (2023):

- PPG Sensor: measures the BVP, allowing to extract the HR.
- 3-axis Accelerometer: motion is measured.
- EDA Sensor: Changes in electrical properties of the skin.
- Infrared Termopile: measures skin temperature.
- Internal Real-Time Clock: high accuracy clock.
- Event button: marks beginnings and ends of events.

Hence, the following signals are possible to be extracted:

- EDA: Electro Dermal Activity measured in microSiemens (μ S)
- HR: Heart rate frequency measured in Hz.
- TEMP: Temperature sensor data measured in Celsius (°C)
- ACC: Accelerometer values in x, y and z directions, measured in g.
- BVP: Blood Volume Pulse.

Dataset Variables

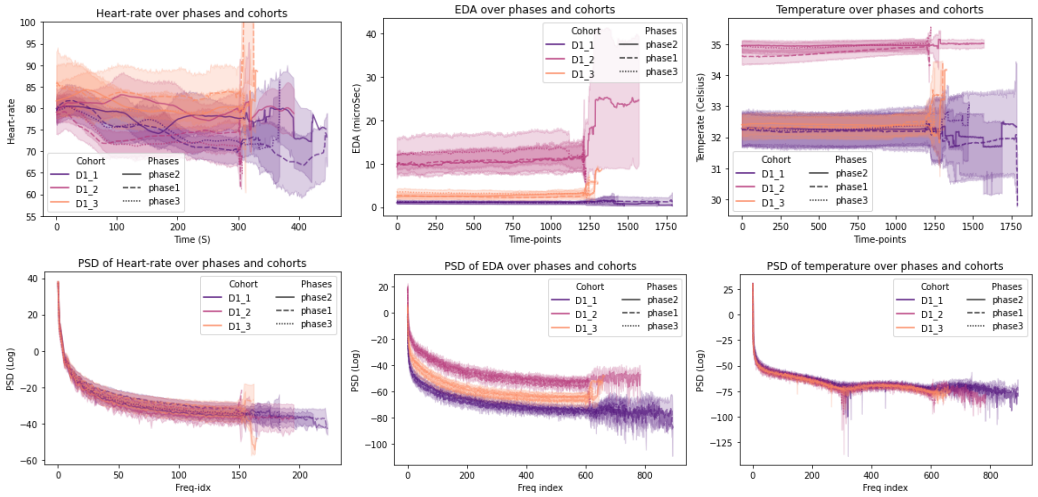
Variables	Description
HR	Time-domain Heart-rate signal
BVP	Time-domain Blood Volume Pulse signal
EDA	Time-domain electro dermal activity
TEMPERATURE	Time-domain temperature signal
ACC	Time-domain accelerometer signal
Round	Puzzling round (1-4)
Phase	Phase of data collection in a round (1-3)
Individual	Index of the individual
Puzzler	Was the individual puzzling or instructing
PANAS responses	self-rated levels of frustrated, upset, hostile, alert, ashamed, inspired, nervous, determined, attentive, afraid, active, difficult
Cohort	Different population groups

Participation: Participants (total $N = 28$) were recruited from our interdisciplinary research group and students and employees within a university department. The ages of the participating women and men ranged from approximately 20 to 42 years. The experiment was run on three separate occasions. D_{11} were completed in the winter and D_{12} and D_{13} were completed in the fall. D_{13} were conducted in four separate sessions; D_{13_1} and D_{13_3} were completed in the morning and D_{13_2} and D_{13_4} were collected in the evening.

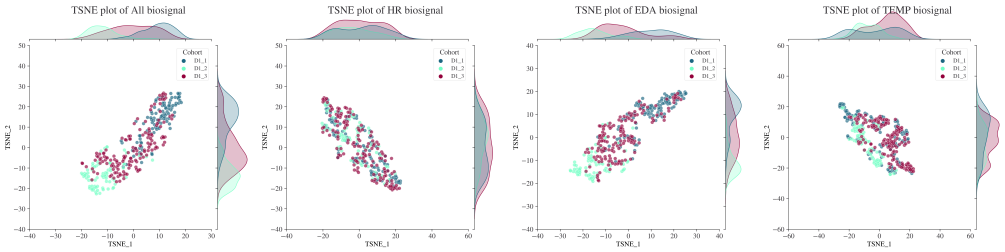
3 TECHNICAL VALIDATION

Signal summary: The mean and standard deviation of the time-domain HR, EDA and temperature signals and their corresponding power spectral densities (PSD) are presented in Fig. 2a with respect to the phases and cohorts. Furthermore, we extracted features using the *NeuroKit2* toolbox and features are visualised using t-SNE and differentiated based on cohort Makowski et al. (2021); Van der Maaten & Hinton (2008). Considerable differences can be observed between the cohorts, despite the replication in the collection methodology.

[2]Due to the non-anonymous nature of the reference, it will be added at the time of the publication.

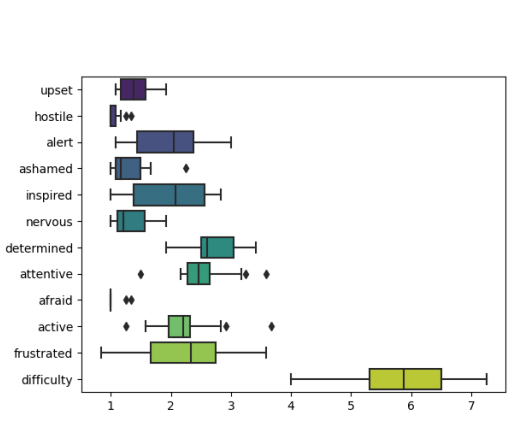


(a) Mean and standard deviation of the time-domain and frequency domain power spectral density (PSD) signals (HR, EDA, Temperature).

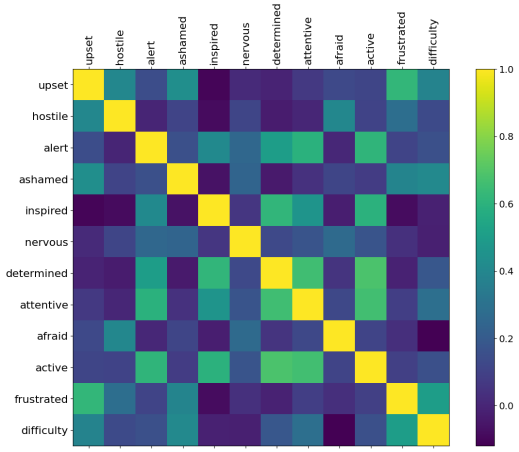


(b) t-SNE visualisation of the features extracted from the signals (all, HR, EDA, TEMP) coloured based on cohorts.

Self-rated Responses: The box plot shown in Fig. 3a presents the distribution of the PANAS dimensions by the participants. The scale for 'frustrated' and 'difficulty' ranges from 0 to 10, while the scale for the rest is from 1 to 5. Responses to *difficulty* is logged by the participants only after the task phase. In Fig. 3b, the correlation between the PANAS variables are shown. Overall, we observe



(a) Boxplots showing the distributions of PANAS variables responses.



(b) Correlation of PANAS variables responses.

Figure 3: Self-rated responses

a positive correlation among the variables, except for some specific cases (alert-hostile, alert-afraid). However, these cases with negative correlation occur with feelings with a very small variance. The

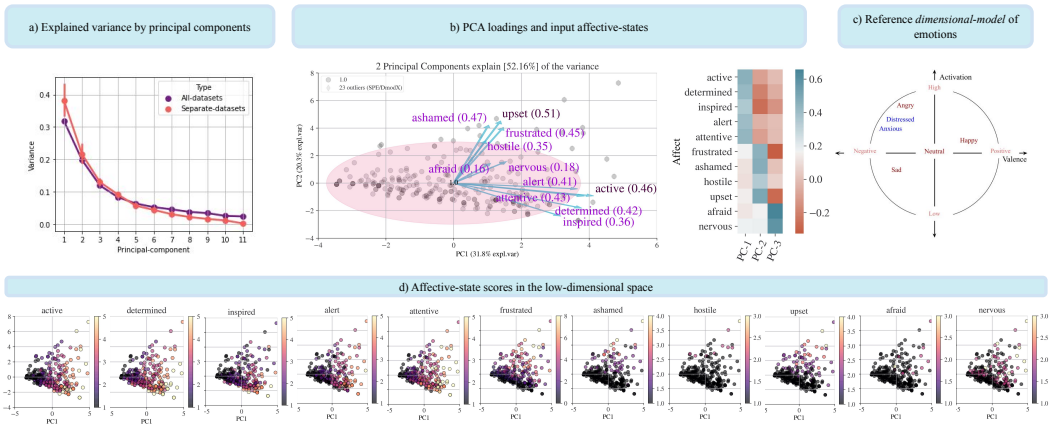


Figure 4: PANAS responses in lower-dimensional space and its correspondence to the dimensional-model of emotions Russell (1980).

highest correlation occurs among positive-positive and negative-negative feelings, such as active-determined, inspired-determined, attentive-determined, frustrated-upset, afraid-hostile.

Phase Analysis: The participants go through 3 different phases (rest, stress, recovery). We group the pre-task and post-task phases and compare the responses to the task phase. From Fig. 1b, we observe that the reported responses are stronger *after* the task phase. Frustration, active, attentive, determined and alert variables are noticeably higher in the task phase than in the rest phase. This is anticipated and further validates the dataset, as participants are likely to experience higher activation during the task phase than during resting periods.

Subspace Analysis of PANAS Responses: We reduced the dimensionality of the PANAS responses using principal component analysis (PCA). The explained variance (a), the loadings with respect to the affective states (b) and the affective-state responses from the participants for each PANAS variable in the PCA subspace (d) are demonstrated in Fig. 4 Wold et al. (1987); we can disentangle and identify 10 emotion variables from the analysis.

4 Conclusion

We present the EmoPairCompete dataset which is one among the few studies on dyads of prosocial behaviours for in-the-wild applications. The dataset comprises of physiological signals: HR, BVP, EDA and TEMPERATURE obtained from the Empatica E4, which is a clinical grade device. The participants are engaged in a team-puzzle task which is organised in phases and rounds. Labels for the affective states are obtained from self-rated questionnaires according to PANAS form. The dataset comprises of three cohorts or collection rounds. Analyses of the dataset shows the possibility of disentangling affective states from the responses and a considerable difference in the signals between cohorts, enabling the development and evaluation of machine learning tools for in-the-wild and high-stakes applications.

References

E4 sensors, 2023. URL `https://www.empatica.com/en-int/research/e4/`.

Andrea Bizzego, Giulio Gabrieli, and Gianluca Esposito. Deep neural networks and transfer learning on a multivariate physiological signal dataset. *Bioengineering*, 8(3):35, 2021.

Giorgos Giannakakis, Dimitris Grigoriadis, Katerina Giannakaki, Olympia Simantiraki, Alexandros Roniotis, and Manolis Tsiknakis. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, 13(1):440–460, 2019.

Jennifer L Hudson and Ronald M Rapee. Parent–child interactions and anxiety disorders: An observational study. *Behaviour research and therapy*, 39(12):1411–1427, 2001.

Dominique Makowski, Tam Pham, Zen J Lau, Jan C Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and SH Annabel Chen. Neurokit2: A python toolbox for neurophysiological signal processing. *Behavior research methods*, pp. 1–8, 2021.

Scott Monteith, Tasha Glenn, John Geddes, Peter C Whybrow, Eric Achtyes, and Michael Bauer. Expectations for artificial intelligence (ai) in psychiatry. *Current Psychiatry Reports*, 24(11): 709–721, 2022.

Kristoffer Vinther Olesen, Nicole Nadine Lønfeldt, Sneha Das, Anne Katrine Pagsberg, and Line Katrine Harder Clemmensen. Predicting obsessive-compulsive disorder events in children and adolescents in the wild using a wearable biosensor (wrist angel): Protocol for the analysis plan of a nonrandomized pilot study. *JMIR research protocols*, 12(1):e48571, 2023.

Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(1):293, 2020.

James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39 (6):1161, 1980.

Muniba Saleem, Craig A Anderson, and Christopher P Barlett. Assessing helping and hurting behaviors through the tangram help/hurt task. *Personality and Social Psychology Bulletin*, 41(10): 1345–1362, 2015.

Philip Schmidt, Attila Reiss, Robert Durichen, and Kristof Van Laerhoven. Wearable-based affect recognition — a review. *sensors*, 19:4079, 2019. ISSN 23719850, 19493045. doi: 10.3390/s19194079.

Karan Sharma, Claudio Castellini, Egon L van den Broek, Alin Albu-Schaeffer, and Friedhelm Schwenker. A dataset of continuous affect annotations and physiological signals for emotion analysis. *Scientific data*, 6(1):196, 2019.

Edmund R Thompson. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (panas). *Journal of cross-cultural psychology*, 38(2):227–242, 2007.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

A Details on Experimental procedure

All participants were given an E4 biosensor and instructed on how to wear it, turn it on and off and mark events. Biosignals were collected under a resting period, an emotion-eliciting condition, and a recovery period as depicted in Figure 1a. After each condition, participants completed a questionnaire. This procedure was repeated four times successively. We instructed participants to turn off the biosensor after each recovery period. Each time the biosensor was turned on, it was allowed to calibrate for one minute. For the resting and recovery periods, we asked participants to find a comfortable position, sitting still and quietly and rest to the best of their ability.

B Similarity Validation

Three stages of experiments are held for the data acquisition process.It is important to understand the similarity of these datasets, and test the results' reproducibility. The analysis of this section directly addresses the second research question *Can we generate reliable and useful biosignal data for the synchrony analysis? Is data consistent along the different experiments?*, which intended to analyse the data consistency along the different experiments. We do this using statistical tests.

There are 2 datasets analysed: original dataset (features) and dataset respect baseline (features standardised after subtracting the rest value of round 1 phase 1). It is important to recall at this point that our dataset does not respect the IID principle, as there are multiple records from the same individual in the experiment. Hence, the use of the statistical tests here needs to be interpreted under that assumption, with illustrative purposes mainly rather than strictly statistical analysis.

For simplicity, a common *p* value threshold of $< .05$ is used for every test. This value shows strong evidence against the null hypothesis H_0, as the probability that the null evidence is correct is lower than 5%. Another very important element to mention is that the rejection of the null hypothesis does not imply that the alternative hypothesis H_1 is true [3] . Hence, other methods, such as visual inspection, are encouraged to reach a conclusion about the datasets' similarity.

CONTINUOUS VARIABLES: ANOVA

The first analysis is the ANOVA, which uses the F-Test to assess how the different experiments differ from each other. Only the continuous variables are taken for the test, therefore the answers from the participants experiments are discarded in this section due to their discrete nature. The distribution of p values on the original and standardized (compared with baseline) datasets are shown in Figure 5

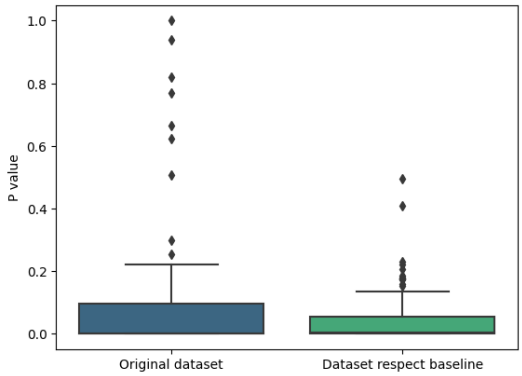


Figure 5: Variables along experiments: Continuous variable validation via the ANOVA test. The box plot shows the low p values obtained in the test, indicating a rejection of the null hypothesis.

The Table 1 presents the % of variables that lie above or below the threshold established of p value, then, accepting or rejecting the null hypothesis.

Dataset	**p< .05**	**p> .05**
Original	70%	30%
Baseline	75%	25%

Table 1: Variables along experiments. ANOVA test *p*-values distribution: Only a 25/30 % of variables accept the null hypothesis. This implies that the continuous variables along datasets do not seem to be similar.

It is observed that the majority of variables, both from the original features and normalised features have p value below the threshold, rejecting the null hypothesis. However, it is important to recall that the ANOVA test show insignificant *p*-values when at least 1 of the datasets compared does not follow the trend of the others.

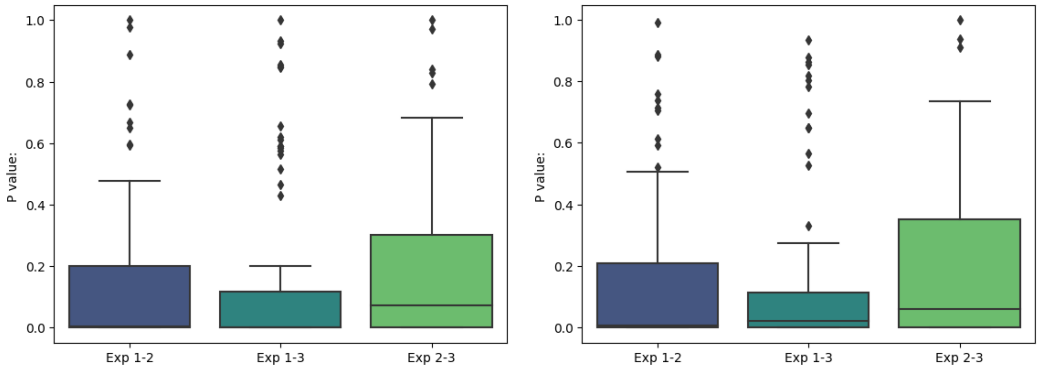
Therefore, the same test must be performed by couples between experiment datasets. The test to be used is not ANOVA, as ANOVA is strictly used for more than 2 datasets. Thus, the reduced version

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017929/

for 2 datasets, the *t*-test, is used.

CONTINUOUS VARIABLES: *t*-TEST

The *t*-test analyses the similarity between 2 univariate datasets. The datasets are compared by pairs between experiment 1,2 and 3, so in total we will find 3 comparisons. The distribution of p values are shown in Figure 6. The notation Exp 1-2 corresponds to the t-test between the values of experiment session 1 and experiment session 2.



(a) Original dataset

(b) Dataset respect baseline

Figure 6: Variables along experiments: Continuous variable validation via T test. Comparing individually the datasets improve the similarity of continuous variables.

Like in ANOVA, it is helpful to identify how many of the variables are found to accept or reject the null hypothesis of experiment session similarity.

Dataset	**Experiment Session**	**p<0.05**	**p>0.05**
Original	Exp 1-2	55%	45%
	Exp 1-3	63%	37%
	Exp 2-3	48%	52%
Baseline	Exp 1-2	63%	37%
	Exp 1-3	63%	37%
	Exp 2-3	49%	51%

Table 2: Variables along experiments. T test p-values distribution comparing the datasets by pairs. Between 37 % and 52 % of continuous variables accept the null hypothesis. These values represent an improvement respect the ANOVA test of previous section.

The results of the T-test show a higher percentage of values above the p value threshold than the ANOVA test. This is because it excludes one of the datasets from the test, which may differ from the other and fail the test.

Experiments 1 and 2, the ones performed before the execution of this thesis, have around a 45% of similarity (percentage above the $p = .05$ threshold), while experiments 1-3 have a slightly lower amount of similarity. However, experiments 2 and 3 show the highest level of variables with p value above the threshold.

DISCRETE VARIABLES: KRUSKAL-WALLIS TEST

A different test is required to do a similar for the set of discrete variables, like the analysis in section B. The use of a different test is explained by the suitability of ANOVA test only for continuous variables and not discrete. For that reason, the Kruskal-Wallis test, which allows discrete variables,

is applied here. The discrete variables analysed are the frustration level, difficulty experienced and the 10 Likert scale feelings stated in the participants questionnaire.

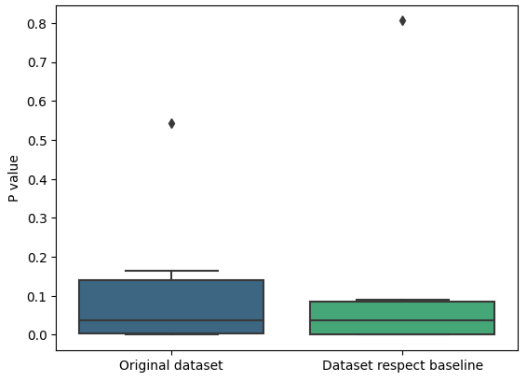


Figure 7: Discrete variable validation: Kruskal-Wallis test

The discrete variables show a higher level of similarity among datasets in respect to the continuous values. There are almost no outliers in the boxplot seen in image 7, suggesting that the distribution of similarities is similar along variables.

As seen in Table 3, the values which do not reject the null hypothesis H_0 correspond to the 50% of the whole dataset. This implies that half of the discrete values mean rank value are equal, while the other 50% have a different mean rank value.

Dataset	**p< .05**	**p> .05**
Original	50%	50%
Baseline	50%	50%

Table 3: Kruskal-Wallis test p-values: Discrete variables

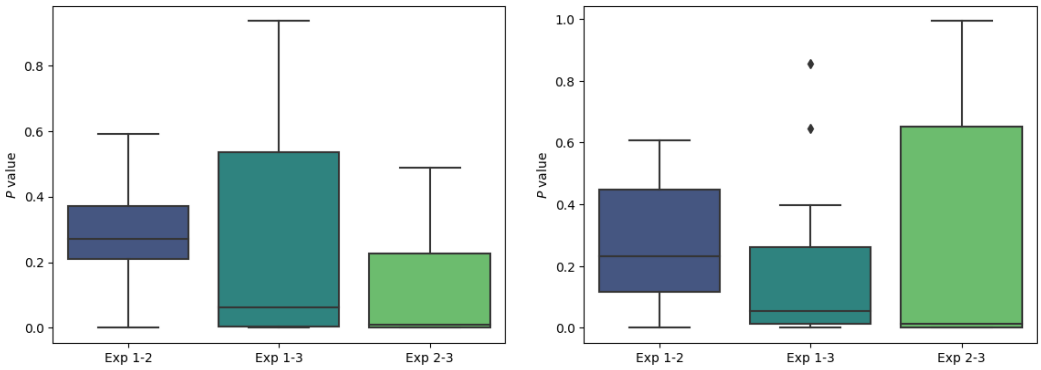
It is important to remind the reader that these discrete variables are completely subjective and dependent on the perception of the participant. Several factors can affect the answers of the participant, and differ the answers among experiment sessions.

DISCRETE VARIABLES: MANN-WHITNEY U TEST

The last statistical test of the section is the Mann-Whitney test, which is equivalent to the reduction of the Kruskal-Wallis test to a comparison of only 2 datasets. Like in the continuous variables, the test is applied by pairs on the 3 different experiment sessions.

Figure 8 shows the distribution of Mann Whitney U test *p*-values for the discrete variables along the experiment pairs. It is noticeable the increase of values above the threshold, which accept the null hypothesis and conclude that the mean rank of the variable along datasets is similar. For the experiments 1 & 2, the median is above 0.2 and most of the values above the p value threshold. The experiment 2 & 3 pair show however a median value lower than the *p* value threshold of .05. This suggests that the majority of discrete variables have rejected the H_0 hypothesis for this specific pair.

Checking the percentages of discrete variables above the threshold from the table above, we can observe a majority of discrete variables with similar mean rank value in the pairs experiment 1-2 and Experiment 1-3, with over 80% of values which accept the null hypothesis. It is interesting to see that, even though the *p* value distribution of Exp 1 & 2 is by average higher than 1 & 3, the % of values above the threshold is the same. This is because it does not matter if a p value is .06 or .6, both of them compute in the same way (as accepting H_0) in the final percentage. The experiment 2 & 3 validation shows very different results from the other 2 pairs. Around 2/3 of the discrete variables rejected the null hypothesis, indicating that only 1/3 of them have similar mean rank in both datasets. This was expected from the boxplot of 8, with a very low median *p* value below .05.



(a) Original dataset

(b) Dataset respect baseline

Figure 8: Discrete variable validation: Mann-Whitney U test along experiments

Dataset	**Experiment Session**	**p< .05**	**p> .05**
Original	Exp 1-2	17%	83%
	Exp 1-3	17%	83%
	Exp 2-3	66%	34%
Baseline	Exp 1-2	17%	83%
	Exp 1-3	17%	83%
	Exp 2-3	66%	34%

Table 4: Mann-Whitney U test p-values: Discrete variables

It is interesting to understand why the participants expressed different feelings in experiment 3 respect to 2. As mentioned before, the level of subjectivity of the responses is very high, and several non-measured factors influence the answers.