PIV++: Evaluating a Personalizable, Inconspicuous Vibrotactile (PIV) Breathing Pacer for In-the-Moment Affect Regulation

Pardis Miri, Emily Jusuf, Andero Uusberg, Horia Margarit, Robert Flory, Katherine Isbister, Keith Marzullo, James J. Gross
Stanford University, University of Maryland, and University of California, Santa Cruz, Intel Labs
parism@stanford.edu, ejusuf@stanford.edu, andero@stanford.edu, andero@stanford.edu,
robert.flory@intel.com, katherine.isbister@ucsc.edu, marzullo@umd.edu, gross@stanford.edu

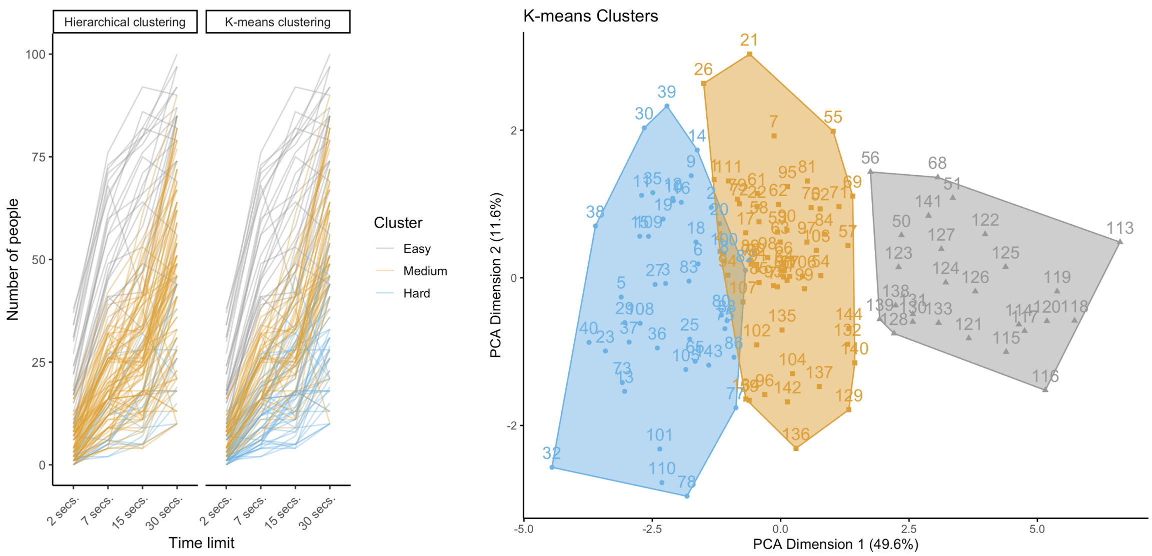

Figure 1. Left: Parallel coordinate plots show the result of both hierarchical and *K*-means clustering. Using Bowden & Jung-Beeman's (2003) data on the number of people who answered each question within 2-, 7-, 15-, and 30-second time limits, questions were clustered into three groups presumably representing difficulty level. Right: Questions in clusters generated by the *K*-means algorithm, visualized in two dimensional-space with axes corresponding to the first and second principal components. Numbers next to points represent question number in Bowden & Jung-Beeman's original study.

SUPPLEMENTAL MATERIAL

Stressor Design

In this section, we discuss our process for ensuring that the two sets of CRA questions had similar levels of difficulty across Stressors 1 and 2.

For each CRA question, Bowden & Jung-Beeman reported the number of people who correctly answered the question within 2, 7, 15, and 30 seconds [3]. We assumed that the fraction of people who correctly answered a question within these time frames was indicative of the question's difficulty level. To group questions by difficulty, we first performed two different clustering algorithms on the answering time data (Figure 1, left) and selected the clustering that had the most visually separable and similarly sized clusters. Then, we used this clustering results to identify pairs of questions with similar difficulty levels.

We compared the clusters created by hierarchical clustering and *K*-means algorithms. First, we performed hierarchical clustering using Euclidean distances between questions and a complete-linkage method. Based on a visual inspection of the resulting dendrogram, we chose to separate clusters at a maximum Euclidean distance between points (dendrogram height) of 8, yielding three clusters of 89, 15, and 25 questions respectively. To determine the appropriate number of clusters to use for *K*-means, we ran the elbow method and the silhouette method. Both the elbow and silhouette plots suggested $k = 3$ as the optimal number of clusters. We ran *K*-means at $k = 3$ multiple times using random initial seeds, selecting the cluster configuration that had the most clearly visually separable clusters when viewed in two dimensions using a PCA. The final clustering had 26, 57, and 46 in the easy, medium, and hard clusters respectively. We chose to use this clustering over the one provided by hierarchical clustering because the clusters were approximately the same size, making it more appropriate for our purpose of grouping questions by difficulty level.

We manually paired together questions of similar difficulty based on the distances between questions in the two-dimensional visualization. For example, questions 50 and 123 were paired together due to their short distance from each other (Figure 1, right). Paired questions were randomly assigned so that one question in each pair would go to Stressor 1 and the other to Stressor 2. This way we ensured that Stressors 1 and 2, while consisting of different CRA questions, would have similar average levels of difficulty.

Overview of Model Selection

In this section we describe how we selected a best model using 26 features and 44 observations from the treatment group.

Before training the models, we split the 44 observations of the dataset into a training set and a test set. The training set was approximately two-thirds of total observations (32 items), while the test set was approximately one-third of total observations (12 items). We balanced the test set so that 6 of 12 observations had a drop in anxiety from Post-stressor 1 to Post-stressor 2, and the other 6 observations either had no change or an increase in anxiety. Balancing was a necessary step to ensure that the trained model's performance could be evaluated on both types of data (drop or no drop in anxiety).

We then trained a small number of XGBoost regression models while optimizing their hyperparameters and reported their averaged performance in terms of goodness of fit on a reserved set of testing data. XGBoost is a Python library that implements a gradient boosted ensemble of decision trees. Gradient boosting uses a gradient descent algorithm to minimize loss with regards to subsequent trees in the ensemble when adding a new tree [4]. In XGBoost, a set of trees are added to the ensemble using an additive training technique to make the final prediction. Additive training involves adding a tree one at a time if doing so helps optimize an objective function—in this case minimizing a loss function of mean squared error—until no more trees can be added to improve the model. In our case, the randomly selected model had 219 trees with maximum depth of 11 and averaged performance of 3.13 of goodness of fit. Optimizing hyperparameters was done using the Hyperopt library in Python, which provides an infrastructure for optimizing hyperparameters of a learning algorithm [2]. For the purpose of model reproducibility, the hyperparameters of our best model were as follows: alpha = 6.0; colsample_bytree = 1.0; eta = 0.25; gamma = 0.7; lambda = 1.0; max_depth = 11; min_child_weight = 2.0; n_estimators = 219, subsample = 0.8; booster = gbtree; eval_metric = rmse.

CHI'20, April 25–30, 2020, Honolulu, HI, USA

DOI: 10.1145/3313831.3376757

Q&A

What values for placement, pattern, and personalization did you use for PIV?

PIV uses a biphasic pattern in which the vibration that indicates inhalation feels different from the vibration that indicates exhalation. The PIV developers tested different settings for the placement of tactors, as well as the order and shape of the vibration pattern [1]. With regards to placement, they found that placing the tactors on the abdomen led to more regular breathing as compared to chest and lower back locations. Order refers to whether the inhalation vibration is more intense than the exhalation vibration or vice versa. The PIV developers found that a more intense exhalation vibration led to slightly less chest movement and a smaller ratio of chest-to-abdomen movement than a more intense inhalation vibration. Finally, shape refers to whether the two types of vibrations differ in their frequency, amplitude, or both. The PIV developers found that a shape with vibrations differing only in amplitude led to participants having more difficulty differentiating between inhalation and exhalation vibrations. Of the remaining two options, they recommended the shape with vibrations differing only in frequency because the corresponding personalization procedure required less time to perform.

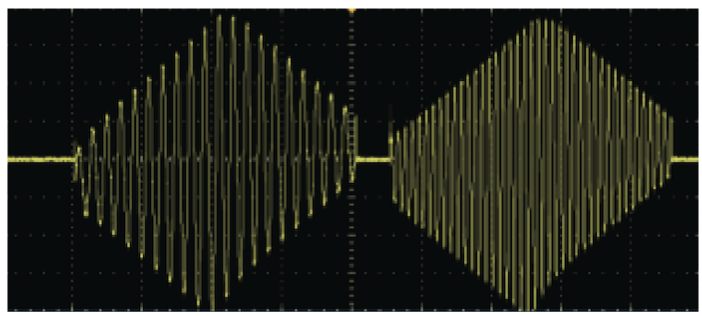

Figure 2. A biphasic breathing pattern with less intense frequency indicating inhalation phase and a more intense frequency indicating exhalation phase (i.e., horizontal strong_exhale). This breathing pattern is captured with an oscilloscope. The x axis represents time and the y axis represents PWM level.

Following the recommendations of the PIV publication for our study, we placed tactors on the abdomen and used a biphasic pattern with a less intense frequency indicating exhalation and a more intense frequency indicating exhalation (see Figure 2). The personalization procedure we used to arrive at appropriate frequencies and amplitudes was as follows: *Step 1: Let me know when you begin to feel the vibrations. Step 2: Let me know when the vibration is easily and vividly noticeable. Step 3A: Let me know as soon as you start feeling the vibration. Step 3B: Let me know when you can no longer feel the vibration. Step 4: Further adjustments. Step 5: Synchronize your breathing.* Steps 4 and 5 were repeated until the participant was comfortable and ready to proceed. For more detail on the personalization procedure, see Supplementary Materials of the PIV publication, "Fine-Tuning Horizontal Pattern" section in Personalization Routine Script.

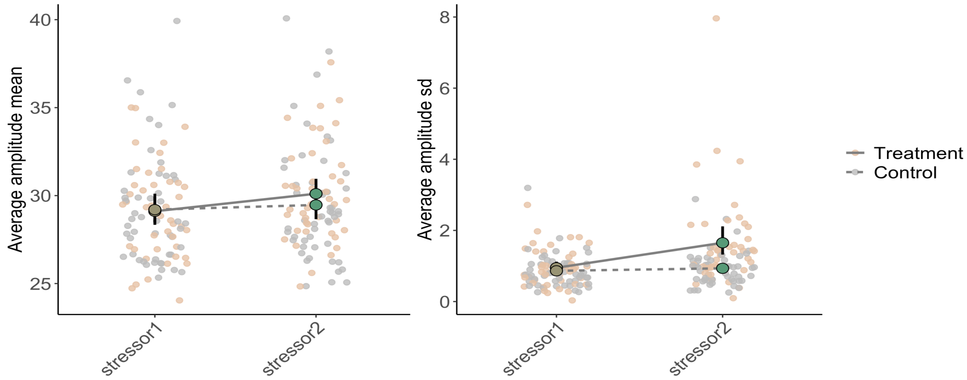

Figure 3. The averaged amplitude mean and standard deviation of chest breathing during stressor 1 and 2 indicate that the pacer indeed influenced the breathing pattern in the treatment group during block2.

Did the pacer intervention change breathing patterns in the treatment group?

Yes. During the pre- and post- stressor 2 stages, both the chest and abdominal breathing patterns of the treatment group were significantly different from the breathing patterns of the control group. This effect is due to lack of a stressor, which allowed for active breathing with the pacer. During the stressor 2 stage, however, the effect was not as large due to the presence of the cognitive stressor, which enforced more passive breathing and switching between passive and active breathing. Figure 3 illustrates the averaged amplitude and standard deviation of chest breathing during the stressor 1 and 2 stages. The interaction effect in the averaged standard deviation between groups and conditions suggests that the treatment group experienced higher depth-wise breathing irregularity. We believe that this is due to switching between passive and active breathing during the cognitive stressor stage. We defer further analysis of physiology data for future work.

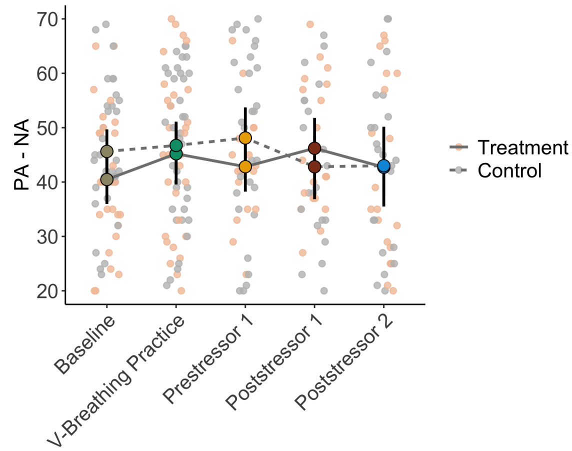

Figure 4. PA - NA

Why did you use the unipolar valence model instead of the valence-arousal model to measure affect?

We had two reasons for adopting the unipolar valence over the valence-arousal model. First, valence and arousal are not two independent measures. There is a V-shaped relation of arousal as a function of valence, i.e. stronger valence in either direction is associated with greater arousal. Second, participants often find positive and negative affect easier to understand in practice than valence and arousal [5].

In accordance with the unipolar valence model, participants were given the following instructions when asked to rate positive and negative affect: *People often report their feelings as a combination of some positive and some negative feelings at once. For example, a young woman who had just eaten a chocolate bar reports a blend of joy and guilt. To report your feelings, please consider using both positive and negative feelings scales.*

You pre-registered studying dependent variables PA + NA and PA NA. You instead reported PA and NA in the paper. Why is that?

In hindsight, we should have pre-registered studying positive affect (PA) and negative affect (NA) rather than their sum and their difference. This is because we collected PA and NA, and given that valance and arousal are not independent measures, studying PA and NA directly makes more sense.

In the paper we reported that we did not observe any interaction effects between condition and group for either PA or NA. Given our pre-registered variables, we also looked for any interaction for either PA + NA and PA − NA. The treatment group saw an increase in PA + NA, but the effect was not significant (intercept = 80.5, β = 0.25, p = .92, CI = $[-4.3, 5.3]$). And, the treatment group saw a decrease in PA − NA, but the effects were not significant (intercept = 17.6, β = -2.87, p = .51, CI = $[-11.2, 6.1]$).

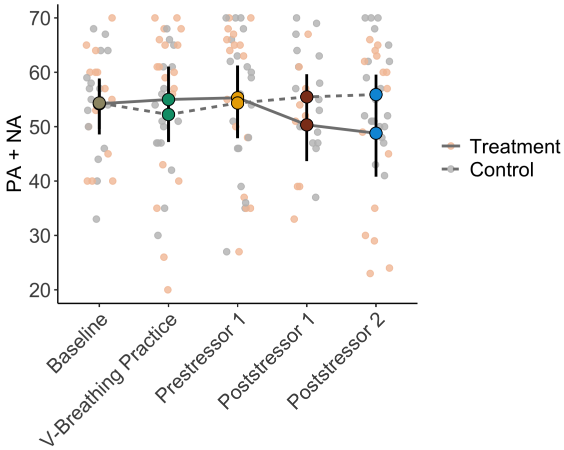

Figure 5. PA + NA

Were treatment and control groups balanced?

Treatment and control groups were found to be balanced across a variety of characteristics. [1]

There was no significant difference between treatment and control groups in years of education (means of 14.43 for treatment group, 14.39 for control group) or area of education (25 STEM, 18 non-STEM and 1 undecided in treatment group; 30 STEM, 20 non-STEM and 2 undecided in control group). Participants were asked on a scale of 0 to 100 how much they had previously undergone training in slow-paced breathing; there was no significant difference between groups (means of 41.9 out of 100 in treatment group, 45.7 out of 100 in control group). Finally, there was no significant difference between treatment and control groups in whether participants were currently playing a wind instrument (0 in treatment group, 2 in control group) or preferred wearing tight or loose clothing (14 preferred tight, 30 preferred loose in treatment group; 20 preferred tight, 32 preferred loose in control group).

Did you check for outliers before running your linear mixed modeling?

We checked for outliers before running a linear mixed model to check for an interaction effect between group and condition (Post-stressor 1 and Post-stressor 2). To check for outliers, we conducted a skewness test on the difference between values at Post-stressor 1 and Post-stressor 2 for each of STAI-6 scores, positive affect, and negative affect. For all three DVs, skewness was found to be within the range of 0.2 to 0.7, which was low enough to suggest the nonexistence of outliers. We did the same process before running a covariance pattern model with unstructured error structure to check for an interaction effect between group and condition (Pre-stressor 1 and Post-stressor 1). Again, we did not find any outliers.

Before running a Linear Mixed model to check for an interaction effect between PIV-user engagement (notice, differentiate, and synchornize) and condition (V-Breathing, Post-stressor 2), we checked for outliers. To check for outliers, we conducted a skewness test on the difference between values at V-Breathing Practice and Post-stressor 2 for each of notice, differentiate, and synchronize. The test identified two participants whose values were over three standard deviations from the mean for at least one measure. These two participants were identified as outliers and excluded from the engagement difficulty analysis.

[1] Chi squared and t-test were employed to test the demographic balance between the two groups.

Collected Self-reported measures during the study

To minimize interaction between RAs and the participants which could induce biases, all questions were presented in the form of questionnaires with a Likert scale or an open ended format. The open ended questions were mostly asked only at the end of the study to ensure equal experiment duration for both Treatment and Control groups until the end of post-stressor 2. The Treatment group were asked more questions regarding the vibrations at the end of the study that contributed to the longer duration (shown in white in Table 1). Table 1 contains a simplified version of questions that were asked at each stage of the study. We randomized the sequence of positive affect and negative affect questions as well as the order of 6 questions in STAI-6 questions.

How do you know the vibrations were not were not distracting and people were actually using the distraction as an affect regulation strategy?

We asked participants how much they found the vibrations distraction and the measure of user of distraction as an affect regulation strategy and the vibrations being distracting were uncorrelated.

Did prior familiarity with CRA tasks have an impact of this study?

This is a potential concern because participants with prior familiarity might remember answers and therefore not find the CRA tasks stressful. The answer to the question, though, is no. At the end of the study, participants were asked whether they had prior familiarity with the CRA creativity tasks. Participants s014 and s073 from the treatment group and participants s028 and s061 from the control group reported prior familiarity. The difference between Post-stressor 1 and Post-stressor 2 STAI-6 scores for these individuals showed no drop in anxiety. In addition, when we excluded these participants from the analysis, we still observed the PIV anxiety reduction effect.

Is this work generalizable to other contexts?

Affect regulation is context specific and effectiveness of the implemented strategy is dependent on numerous factors including culture and individual differences. It is therefore an open question whether the tools described in this report will generalize to other populations and contexts, and this is an important question for future research.

Table 1. List of during-the-study questions for treatment group (in white) and for both treatment and control (in gray).

	Randomized order	(1) Baseline	(2) V-breathing practice	(3,4) Pre- & Post-stressor 1	(5) Post- stressor 2
How positive do you feel right now? How negative do you feel right now?	⧢	●	●	●	●
If your negative feelings were related to the vibrations, could you describe how?			●		○
STAI-Y1 6	⧢	●	●	●	●
How difficult was it to notice the vibration pattern? How difficult was it to differentiate the inhalation phase from the exhalation phase in the vibration pattern? How difficult was it to synchronize your breathing with the vibration patterns?	⧢		●		○
Did you feel anxious during the creativity tasks? What were the activities you engaged in to lower your anxiety before, during, and after the first and second set of creativity tasks? Here are various strategies people use to lower their anxiety. Please indicate the level of engagement and success in the following activities. 1. I tried to think in a way that helped me stay calm 2. I tried to distract myself 3. I tried to suppress my anxiety 4. I tried slow-paced breathing 5. Other (please specify): Do you have prior experience with similar creativity tasks before participating in this study?					●
What do you think was the purpose of the vibrations? 1. To pace my breathing 2. To distract me 3. To help me focus 4. To annoy me 5. Others (please specify): During the creativity tasks, what percentage of the time would you say that you were able to synchronize your breathing with the vibrations? If you were not able to fully synchronize your breathing with the vibrations, what would you say were the barriers? How much did the vibrations had an impact on you? If the vibrations had an impact on you, describe the impact in more detail please. To what extent during the study did you feel like you wanted to turn off the vibrations? If you found yourself wanting to turn the vibrations off, walk us through your experience. To what extent did you feel that the vibrations were distracting? If you found the vibrations distracting, walk us through your experience. To what extent do you think the vibrations affected your performance? If the vibrations affected your performance, describe in more detail.					○

ACKNOWLEDGEMENTS

We would like to thank our dedicated research assistants Alyssa Jackson, Anna Speder, Eman Magzoub, Elaheh Salehi, and Satvir Basran for their time and commitment to this project. We also would like to thank professors Dan Yamis, Tobi Gerstenberg, Mike Frank, Heather Culbertson, Eun Kyoung Choe, Sally Olderbak, Amanda Lazar, Niklas Elmqvist, Richard Harvey, and Erik Peper for their time, support, and guidance. Lastly we would like to thank graduate students Maia ten Brink, Mihir Mongia, Jonas Schone, and our former lab manager Julia Ryan for their unconditional support.

REFERENCES

[1] Authors Left Anonymous. 2019. PIV: Placement, pattern, and personalization of an inconspicuous vibrotactile breathing pacer. (2019).

[2] James Bergstra, Dan Yamins, and David D Cox. 2013. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in science conference*. Citeseer, 13–20.

[3] Edward M Bowden and Mark Jung-Beeman. 2003. Normative data for 144 compound remote associate problems. *Behavior Research Methods, Instruments, & Computers* 35, 4 (2003), 634–639.

[4] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 785–794.

[5] Peter Kuppens, Francis Tuerlinckx, James A Russell, and Lisa Feldman Barrett. 2013. The relation between valence and arousal in subjective experience. *Psychological Bulletin* 139, 4 (2013), 917.