Primary Subcommittee Selection

Understanding People: Theory, Concepts, Methods

Secondary Subcommittee Selection

Health

Orange: **1. Venue Suitability**
Green: **2. Method and Collected Measures**
Purple: **3. Data Analysis**
Blue: **4. Accessibility to supplementary material**
Yellow: **5. Misc**
Gray: Not addressing
Brown: my rebuttal

Thank you for the thoughtful reviews. We refer to the submission as PIV++ and the paper that discusses the design of the device as PIV. We identified 5 key areas of concerns and provide summaries of our responses for each.

1. Venue Suitability

Reviewers had concerns about whether this should be a paper, its impact, the need to discuss intervention type, and the rationale for exclusion of physiological analyses.

As R1 noted, this report is part of a larger project. The present report will include one physiological outcome, namely a simple manipulation check (the average amplitude of breathing), but its main focus is evaluating a promising anxiety-reducing technology with a focus on self-reported anxiety. This focus was explicitly defined in the pre-registration, and we believe that maintaining this focus preserves the quality and credibility of this report.The secondary contribution includes the link between the user-PIV engagement and the stressor and the novel use of Shapley values to determine the link between individual differences and potential benefits from PIV.

Our rationale for electing to report these findings via conference proceedings is that there is growing interest in using technology to assist with affect regulation, and we have a sense of urgency to provide these methodological starting points to others in the field. There are many pressing clinical applications. For example, we are currently investigating the efficacy of PIV for youth diagnosed with autism (ASD) inspired by PIV++ findings. We believe that sharing the present findings with the community via conference proceedings will allow for much more rapid dissemination of findings. At the same time, we will release the dataset, the data analysis code, and the materials needed to reproduce our study. This will preserve nuances that may get lost in a conference proceeding format.

Our rationale for leaving the other physiological measures for a subsequent report is that we anticipated (and found) that the reports of anxiety experience were not correlated with the physiological indicators. Given that our primary focus here is on anxiety experience, and that we are currently at the limit of what we can do in a conference proceeding, we believe that the best path forward is to treat the physiological measures in a separate, longer report.

- *[AC2] I do not think that it should be a conference paper. Instead, the authors should consider revising this manuscript and submitting this to TOCHI instead, **as an extension to the implementation of the system**. The thoughtful design of systematically evaluating this implementation has implied nuances that are currently lost within the format of a conference paper.*

- *[AC2] **better suited as a journal article**. A conference paper sells this short because it does not allow the authors to expand on the theoretical motivations for affective regulation and to address why a systematic evaluation such as this is necessary (and currently lacking in the growing field of such applications). In its current form, I doubt that it will have long-lasting impact in spite of the efforts clearly taken by the researchers to perform a robust and replicable evaluation.*

- *[AC2] **# Design implications** Without having access to the prior work on PIV, I can only say that the design implications sub-section is well-written and represent reasonable generalizations given the results. The authors also highlight the limitations of their interpretations, for example with regards to the relative effectiveness of **implicit and explicit** involvement technologies. **I feel that this aspect ought to be expanded upon** and, if so, will have a wider impact than currently.*

- *[R1] **Dense and more suited as a journal** but,I look forward to seeing this paper presented in CHI this year.*

- *[R1] I appreciate the way the authors presented the information in the paper in 10 pages, although **this study is clearly a building block in a bigger project which would be perfect for a journal publication** (especially that the supplementary material is already the size of another paper)**.***

- *[R1] why did the authors use **multiple other sensors during the study**: EDA/Breathing Gauge/Pulse and temperature? Was the data from these sensors used in any way to confirm the authors' hypothesis or for general analysis of the stressor task?*

- *[R1] Was there a difference between those who **scored** well and those who didn't with respect to their responses to the questions?*

- *[R2] **The paper is well written** -- albeit some verbosity, for example in the **introduction or in the discussions***

- *[R2] **my score is low because [lack of physio analysis]:** there is no data concerning the actual breathing patterns of participants. As stated late in the conclusions, the analysis of breathing measurements would unveil key elements for the validation of the system, as it would help to determine up to which point users synchronized their breathing with the haptic feedback. Authors did not write anything about that, and yet as per the experimental protocol they did record physiological signals, and **yet as per the TOCHI paper they do have the entire signal processing pipeline ready**.*

- ***[R2] Author should release their dataset**, opening new opportunities for researchers interested in such a topic.*

- ***[R2] Hence I cannot fathom why such results are not part of this paper. This should not be relegated to future work.*** *I would rather remove the entire section about the model instead; better to have* ***solid conclusions*** *about PIV before investigating further interaction with participants' traits and states.*

- *[AC2]****# Title*** *consider reporting the findings or the research question instead.* ***Being explicit about the research contribution*** *should increase perceived impact.*

- *[AC2]* ***# Introduction:*** *explicit in reporting their findings right at the beginning.* ***clearly state their findings and to position these findings in light of other work.*** *This is because affect regulation technology is a semi-mature field and it is, thus,* ***suitable for the readers to expect a consolidation of main findings across various attempts*** *(also see, comments on 'Discussion'].*

- *[AC2] There is an overemphasis on the pre-registration. While I am personally in favor of Open Science, your practice is currently phrased in a way that appears to be irrelevant to the scientific work per se; please note that the original statement released by CHI for pre-registration is no longer actively pursued. Hence,* ***a footnote*** *would suffice unless the authors feel that there is a stronger point to be made, namely how* ***pre-registration is necessary in order to lend strength to their current findings****.*

2. Method and Collected Measures

Reviewers requested clarifications about qualitative data, experiment length, and user engagement measures.

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 asked only at the end of the study to ensure equal experiment duration for both T and C groups until the end of post-stressor 2. The T group were asked more questions regarding the vibrations at the end of the study that contributed to the longer duration. We will add a table with questions that were asked at each stage of the study.

- *R1] Did the authors collect any* ***qualitative data*** *that is not in questionnaire form? It would have been nice to know the actual feedback from the users on the effect of the stressor task and their "perceived" feelings of anxiety before and after the stressor and breathing, especially that they were shown their scores during the task.*

- *[R2] Why does the length of the experiment vary to much between participants (60 to 90m)?*

- *[AC2]****# Results and Discussion:*** *I found the exploratory analyses frustrating to review. For "Stressor and PIV-User Engagement", it was difficult for me to readily understand the differences between the levels of Engagement-type.* ***I reviewed the Methods section***

again but could not readily establish a link between this variable and the methods. *I feel that this section presents some interesting findings but was not able to easily understand what the authors' intended to communicate within the space afforded here.*

3. Data Analysis

R2 raised concerns about p value threshold, post hoc significance, model cross validation and overfitting.

To clarify, we didn't use a Bayesian approach, but bootstrapped the effect CI calculation which suggests the existence of an effect. We will report that the found effect is not post-hoc significant. Note, use of personalized BPM as a confounding variable didn't enhance the model fit.

We also wish to note that:

1 xgboost is a regressor and we did not use binary labels in the model.
2 We didn't use k-fold. We used StratifiedShuffleSplit with n_split = 1 and random_state instead. In theory, depending on a dataset, use of one random split (balanced with label classes) may or may not change the accuracy of a model prediction. We will use n_split > 1 and report the averaged accuracy instead.
3 The parameters of the xgboost model were selected in a cross validation fashion (i.e., fit on a training set and the performance was reported on a held out testing set) which controls for overfitting.

Chi squared and t-test were employed to test the demographic balance between the T and C groups.

- *[R2] Compared to the robustness of the stats I was slightly disappointed to see that a threshold of 0.05 was chosen for significance, a value might be too high to ensure strong conclusions [a]. That, combined with the fact that authors might not have corrected p-values for multiple comparisons could explain some discrepancies authors pointed out, as for example about interaction with Distraction in the Model.*

- *[R2] Page 3, authors mention no significant difference between the two groups, without stating which tests were employed.*

- *[R2] While studying the model, why did authors chose to transform variable to binary categories? I might miss a point here, but* ***regression models*** *could be used to find relationship between continuous values and observed variable, with results that might easier to interpret and put into practice by other researchers.*

- *[R2]* ***Speaking about the model, were the training and testing sets fixed? Without cross-fold validation I fear there might be over-fitting, and it would be hard to test the reliability of the predictions, diminishing the importance of the section.***

- *[R2] Authors submitted Q&A and various additional details as supplementary materials. This practice is unusual to me, and while it could be a good and handy one, it feels a bit like cheating, the actual page count going beyond well 10 (+ new references!). For instance important details about model training and testing are omitted in the main paper, in particular about training and testing sets.*

- *[R2] In the meditation stage, was the slow-pace breathing fixed or tuned depending on participants breathing rate baseline? Did author investigating confounding factor regarding the actual breathing rate of participants?*

4. Supplemental material

AC2 raised concerns about lack of clarity of the exploratory analysis, lack of illustration of haptic vibration pattern, and vagueness in use of PIV recommendations in PIV++ and the ML model.

We did upload the anonymized PIV paper and a Q&A for PIV++ in the supplemental material of this submission. The Q&A document was meant to provide 1) a short summary of PIV along with a figure of haptic shape and explanation of what recommended analysis we were referring to in PIV++; 2) a short background review on xgboost regressor and hyperopt package, and further explanation of the model for those interested in reproducibility of our work.

- *[AC2] If this work is accepted, the authors should focus primarily on rewriting the exploratory analyses to fit better with the methods, design implications, and conclusions. They appear to have been written in haste, without a careful integration into the global narrative. Thus, one can only take the authors' interpretations at face-value without having the opportunity to evaluate the analytical process critically.*

- ***[AC2] My main recommendation is for the authors to consider illustrating the vibrotactile pattern of PIV as a time-series to contrast this with EmotionCheck and/or Doppel. This would be a more intuitive description than p3 col1 par1.***

- *[AC2] The motivation, namely "We have used their recommendations in our research to conduct a thorough evaluation of PIV.", is vague. It will be hard for the reader to determine what constitutes a "thorough evaluation". I advise the authors to paraphrase this in terms of research questions.*

- *[AC2] Instead, this report currently serves as service evaluation of an existing implementation, which the reviewers are unable to access given the double blind procedure. Having access to the tailored evaluation without a clear link to the design of implementation itself is a wasted opportunity.*

- *[AC2] Similar challenges to the reader were experienced when I read the subsection on the predictive model of the efficacy of this intervention given individual differences. Once again, I feel that the authors did interesting work here. [AC2]However, I'm not able to fully appreciate the work or to evaluate it critically without an in-depth **explanation of the models and critical features**. The way that it is currently reported, I can only accept the interpretation of the researchers at face-value. It should also be noted that I am not familiar with the classification approach adopted by the researchers and, hence, could require more information or interaction with the reviewers in order to effectively evaluate this work.*

5. Misc

Reviewers requested clarifications on PIV efficacy outside this study context and the feature selection criteria for the ML model.

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.

We deployed credible and widely adopted questionnaires that capture constellations of personality and affect regulation traits across various cultures including Big 5 and Emotion Regulation Questionnaire. This was the criteria for selection of before-the-study features. The criteria for selection of during-the-study features was primarily user-PIV interaction experiences and state affect regulation strategies deployed during the study.

- *[R2] The readability of various figures should be improved, e.g. Figure 3 can present bigger plots; it is next to impossible to read any values in Figure 6*
- *[R2] PIV is not that inconspicuous if it necessitates **noise cancellation to operate** and various elements to be strapped on the body. More iterations are needed for the design as much as for the engineering.*
- *[R2] I am unsure up to which point it is possible to directly c**ompare the different scores related to user engagement and state that (perceived) synchronization could be more difficult than noticing vibrations. Up to which point these various scales are normalized with one another?***
- *[R2] **Could author comment about the effectiveness of PIV** -- and about reputability of the results -- with different stressors, or in different contexts?*
- *[R2] Up to which point authors decided on some criteria before studying some of the features selected by their model?*

Not addressed

- *[R1] The author mentioned that the PIV could be used by professional therapists to be prescribed to people suffering from anxiety or other mental health issues, but definitely more situations may exist. Reflecting on them in the grand scope of HCI research would be useful for future work.*
- *[R2] The paper might be shorter and easier to follow if the (many) results were presented in a synthetic table*

- *[AC2] This manuscript was written **somewhat hastily** (perhaps due to the conference deadline) and a reiterative process with peer reviewers would help to iron out some avoidable oversights or sub-optimal phrasing.*