Author response to reviews of

Manuscript PRPF-D-20-00087

The emotion–facial expression link: Evidence from human and automatic expression recognition

submitted to *Psychological Research*

RC: ***Reviewer Comment*** AR: Author Response □ Manuscript text

1. Editor Comments

RC: **Dear Dr Tcherkassof,**

We have received the reports from our advisers on your manuscript, "The emotion–facial expression link: Evidence from human and automatic expression recognition", which you submitted to Psychological Research.

Based on the advice received, I feel that your manuscript could be reconsidered for publication should you be prepared to incorporate major revisions. YOU ARE KINDLY REQUESTED TO ALSO CHECK THE WEBSITE FOR POSSIBLE REVIEWER ATTACHMENTS!

In order to submit your revised manuscript, please access the journal's website.

We look forward to receiving your revised manuscript before 25 Aug 2020.

With kind regards, Ruth Krebs Associate Editor Psychological Research

AR: Dear Dr. Ruth Krebs,

We would like thank you for taking the time to consider our manuscript for publication at *Psychological Research*, and the opportunity to resubmit a revised copy of this manuscript. We would also like to take this opportunity to express our thanks to the reviewers for the positive feedback and helpful comments for correction.

We believe such changes have resulted in an improved revised manuscript, which you will find uploaded alongside this document. The manuscript has been revised to address the reviewer comments, which are appended alongside our responses to this letter.

In the following we address the respect points made by both the editor and the reviewers.

2. Reviewer #1

2.1. Main Points

RC: The work of this article gives new insights about the connection of emotional self-reports and observer categorizations and automatic recognition of facial movements. The data are very interesting and this work should be published. But off course some things could be made better in the paper. The most important things are mainly in the methods and results section.

AR: We greatly appreciate the Reviewer's positive comments. We have carefully revised the manuscript according to the Reviewer's insightful comments and provided point-by-point responses as follows.

2.2. Method section

RC: Page 15, line 57: in 2-3 or more lines should be described which kind of emotion inductions have been used for the data base.

AR: As requested a short presentation of emotion elicitation task has been added as follow:

> Two different types of task were used to elicit emotions. Some tasks were passive and consisted in the participant watching videos selected to trigger emotions (e.g. television commercials). Other tasks were active; they required the participant to interact with a computer (e.g. to answer questions or to test a flawed software).

RC: Page 16 line 39: why did the annotators only rate 232 of the 358 videos? And how have been the videos selected? Randomized selection or other criteria? Which? Why is not every video rated the same time? This especially is a point which looks not understandable.

AR: The annotation process was quite complicated. Because the total duration of the database is XXX min, we could not ask each participant to annotate it entirely. In addition we had to insure that each recording is annotated by at least 20 participants to obtain an sufficient agreement. Therefore, we have randomly selected a batch of 10% of the recordings to be annotated until all the videos are annotated at least 20 times. Then, another batch of 10% of the database was randomly selected. However, to avoid their cognitive fatigue, we have limited participant's experiement duration to 30min. In order to add this element and to make it clearer, the annotation procedure has been completely rewritten as follow:

> To analyse how the recorded facial expressions are perceived, 1383 participants were recruited to watch and to annotate these recordings. An iterative procedure was setup to insure that each recording has been annotated by at least 20 participants: 10% of the video database were randomly selected and annotated until all the videos reached the amount of annotation required, then another section of the database is randomly selected among the remaining recordings to be annotated. To avoid potential decrease in the quality of the annotations due to cognitive fatigue, each participant only annotated recordings during a maximum of 30min. As a results only 232 out of the 358 videos have been annotated. Each video was annotated 29 times on average (*SD* = 12).

2.3. Results section

RC: Page 19 line 17 to 20: Which kind of correlation was computed here? A pearson correlation? It should be a point-biserial correlation as one of the two variables is dichotomous.

AR: It appears that point-biserial correlation for two dichotomous variables is mathematically equivalent to calculating the Pearson's product-moment correlation and produces the same results (Bass & Ager, 1991).

The following reproducible example in R shows that results are identical between point-biserial correlation using the function `biserial.cor()` from the package {ltm} and Pearson's product-moment correlation using the function `cor.test()` from the package {stats}:

```r
x <- c(0, 1, 0, 1, 0, 1, 0, 1, 0 , 1)
y <- c(0, 0, 0, 0, 0, 1, 1, 1, 1 , 1)

ltm::biserial.cor(x, y, level = 2)
```

```
## [1] 0.2
```

```r
cor.test(x, y, method = "pearson", alternative = "two.sided")$estimate
```

```
## cor
## 0.2
```

The following example tests the computation of the point-biserial correlation and the Pearson's product-moment correlation with our data (i.e. correlation between participant's self-report and observers' annotation):

```r
recognition_result %$%
  ltm::biserial.cor(sr_score, hr_score, level = 2)
```

```
## [1] 0.2403944
```

```r
recognition_result %$%
  cor.test(sr_score, hr_score, method = "pearson", alternative = "two.sided") %>%
  use_series("estimate")
```

```
##       cor
## 0.2403944
```

As the two functions are providing the exact same result we can insure the integrity of our results to the reviewers using the Pearson's product-moment correlation. However, as our previous manuscript only indicate Pearson's correlation, the new manuscript now mentions Pearson's product-moment correlation in order to disambiguate this point.

RC: Page 21 line 32: "expression intensity", where are expression intensity scores? Could you please bring the data into the paper?

AR: In the previous version of the manuscript, “expression intensity” was referring to a result usually obtained in facial expression recognition studies (see the recent Krumhuber, Küster, Namba, & Skora, 2020 for example). However this assumption is just an interpretation and is not supported by data. In order to correct this point, the sentence has been rewritten as follow:

> This difference between emotions is usually observed in the literature (e.g. Krumhuber et al., 2020). Multiple reasons can explain why *happiness* and *disgust* are more easily recognized than *anger*, *surprise* and *sadness*. It is possible that these expressions are usually more intense than others. In addition, *anger* and *sadness* as non-socially desirable emotions may be have been felt but not expressed.

RC: Page 21 line 40ff: Why should undetermined emotional states reveal that a 6-point Likert scale has a limit? What kind of limit? What would be better to do in future? Please explain in the text.

AR: Many thanks, indeed this point requires some explanations. “undetermined” is used when more than one label obtain the highest score. As there is a tie between two or more label, it is not possible to identify a dominant one. In term of affective states, two emotions can be felt simultaneously with the same intensity. However, if one emotions is felt with a lower intensity, it is possible that likert scales are not discriminant enough with 6-point and results in an “undetermined” labels even if the two dominant emotions are not felt with the same intensity. To better discriminate between labels, an alternative is to use a continuous slider from 0 to 100. This point has been added to the new version of the manuscript as follow:

> “Indeed, Likert scales may not be able to discriminate between two dominant emotion felt with different intensities due to their reduced number of possibilities. To better discriminate between emotions, an alternative would be to use a continuous slider from 0 to 100 (see for example Lottridge & Chignell, 2009).”

2.4. Minor points

RC: Page 4 Line 44/45: one would not assume total agreement among all people based on basic emotion theory, so "totally" should be changed by another term like "mainly" or something similar.

AR: The end of the sentence has been rewritten as follow:

> “… and why observers can make mistakes in recognizing facial expressions of basic emotions, among others.”

RC: Page 5 Line 16/17: Isn´t the point of Kraut and Jonston that according Basic Emotion Theory the smile is only a sign of expression of joy? So I would write: "that a smile is only the major".

AR: The sentence has been changed as suggested.

RC: Page 5 line 58 ff: maybe smiling is not the right sign to be looked directly after winning a contest. Triumpf is usually exposed immediately after the contest and other emotions like joy or pride come a little later (see the work of Matsumoto and colleagues on that). So this argument from Crivelli et al. (2015) does not hold in that context.

AR: To see with Anna

RC: Page 18: line 56: "two-sided pearson correlation" probably means two-tailed t-tests of pearson-correlations, or not?

AR: Many thanks for pointing our mistake, it is indeed a “two-tailed t-tests of pearson-correlations” and it has been changed in the manuscript.

\RC{Page 20 Line25: “60% and 80% accuracy”; please give a citation.}

This information has been precised with a reference to Krumhuber et al. (2020) as follow:

> “These results are far from those classically obtained in the literature for emotional facial expression recognition (Krumhuber et al., 2020, report an average accuracy of 65%).”

RC: Page 23 Line 56: what is correlated here (r=.22)? A correlation of two accuracies does not make sense in my mind. If yes, what are the units? Emotion categories, events or. . . ???

AR: Our formulation was indeed incomplete. For each of the video and emotion categories, all three of self-report, human observers and automatic recognition are encoded as 1 for the emotion category recognised and 0 otherwise. Therefore the correlation corresponds to the correlation between human observers and automatic classifier recognition of emotion categories. This element has been added to the manuscript as follow:

> “As previously mentioned, human observers appear to be more accurate than the automatic classifier to recognize an individual’s subjective feeling (human observers Accuracy = 0.43; automatic classifier Accuracy = 0.30). However, both make similar mistakes as the two-tailed t-tests of Pearson’s product-moment correlation between human observers and automatic classifier recognition of emotion categories is significant ($r = .22$, 95% CI [.17, .27], $t(1384) = 8.31$, $p < .001$).”

RC: Page 31 line 9ff: I do not understand the point under secondly, automatic classifiers assume for every person unique facial expression for one emotion?? Did I understand that right? Could you explain that more in detail. So maybe in the methods section a more detailled description of the algorithm and how it works, would help. . . ; as this is a psychological journal a principal description of the characteristics of the algorithm would help (that does not mean to explai

AR: The output of the automatic classifier is a multivariate timeseries matrix with, for every labels, a probability of expressing the corresponding label for each video frame. Therefore, the equation used to identify which label corresponds the most to the overall video is crucial. In order to make that point clearer, the following sentences have been added to the method section:

> “For each video frame, Affdex identifies the probability $p_{video_i.label_j.t_k}$ of expressing each of the six basic emotion labels (*i.e.*, *anger*, *disgust*, *fear*, *happiness*, *surprise* and *sadness*) as well as additional psychological states such as *valence*, *engagement* or *contempt*, and facial features such as *cheek raise*, *eye widen* or *jaw drop*. The result of Affdex is a mutivariate timeseries output which provides for each label a probability from 0 to 100 (rescaled from 0 to 1 for the analysis) of expressing the corresponding label. For this study only the six basic emotion labels are analysed to match with self-reports and human observers.”

3. Reviewer #2

RC: 1. The paper did not distinguish clearly between opinion and empirical evidence. Please revise.

AR: See with Anna

RC: 2. The paper’s arguments must be built on an appropriate base of theory, concepts, or other ideas. It is recommended that the authors refer to Eigen Face Approach.

AR: The Eigen Face Approach is used in the development of the automatic classifier’s algorithm to maximise emotion classification accuracy (Cendrillon & Lovell, 2000; Turk & Pentland, 1991). Here, the automatic classifier has already been developed and our team is not involved in Affdex’s development. Consequently, it appears that introducing the theory and concepts behind the Eigen Face Approach is not relevant in the context of this study.

RC: 3. The authors did not make clear emotion taxonomy. Ekman’s basic set of emotions and Russell’s circumflex model of affect are recommended to elaborate.

AR: See with Anna

RC: 4. What is the significance of this study? must be explained with a sensible reason.

AR: See with Anna

RC: 5. The paper must contribute to a critical understanding of the issues.

AR: See with Anna

RC: 6. You’re highly advised to show yourselves as authors while explaining theme of research in the paper. Keep the available references and at the same time let’s the readers identify your own ideas and voice concerning what you cite in Introduction.

AR: See with Anna

RC: 7. The selection of the participants, to me, is vague.

AR: In order to precise the procedure used for the selection of participants, the following sentences have been added to the “Emotion Elicitation” section, and to the “Human Facial Expression Recognition” section:

> “For the emotion elicitation experiment, 358 encoding participants (182 females, 176 males, M_{age} = 47.9, SD_{age} = 9.2) were recruited to perform one out of 11 emotion elicitation tasks designed to trigger a positive, a specific negative or a neutral emotional state. These participants were recruited by a private company for a study supposedly devoted to an”ergonomic visual task" (cover story). After a description of the general aims, participants agreed and signed the experiment consent form. At the end of the elicitation task, they received an equivalent of €50 in voucher for their participation. A second consent form was signed by the participants to allow their video to be processed for research purposes."

> “To analyse how the recorded facial expressions are perceived, 1383 participants were recruited among social science under- and post-graduates through advertising to watch and to annotate these recordings. They received a course credit for their participation. After a description of the general aims, participants agreed and signed the experiment consent form.”

RC: **8. I wonder why the authors did not implement Principal Component Analysis.**

AR: PCA are used to infer one or multiple latent constructs from a series of variables. Here, the purpose of this paper is to investigate the relationship between self-reported emotions and emotion recognised by human observers and by an automatic classifier. Consequently, an approach based on correlations appeared to be more relevant than a PCA.

RC: **9. Discussion needs a deeper look. The plausible reasons behind obtaining these results are not satisfactory. Moreover, the author/s did not discuss the validity and accuracy of the data. They did not state what their study adds to the body of literature. The last point is that in the discussion, the authors must compare/contrast their results with similar studies, mentioned in literature.**

AR: See with Anna

RC: **10. It is recommended that authors consider APA, 7th edition in in-text citations.**

AR: As requested the in-text citations now uses APA, 7th edition reference style.

4. References

Bass, A., & Ager, J. (1991). Correcting point-biserial turnover correlations for comparative analysis. *Journal of Applied Psychology*, *76*(4), 595–598.

Cendrillon, R., & Lovell, B. (2000). Real-time face recognition using eigenfaces. In *Visual communications and image processing* (Vol. 4067, pp. 269–276). International Society for Optics; Photonics.

Krumhuber, E. G., Küster, D., Namba, S., & Skora, L. (2020). Human and machine validation of 14 databases of dynamic facial expressions. *Behavior Research Methods*, 1–16. `https://doi.org/https://doi.org/10.3758/s13428-020-01443-y`

Lottridge, D., & Chignell, M. (2009). Emotional bandwidth: Information theory analysis of affective response ratings using a continuous slider. In *IFIP conference on human-computer interaction* (pp. 111–114). Springer.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71–86.

Bass, A., & Ager, J. (1991). Correcting point-biserial turnover correlations for comparative analysis. *Journal of Applied Psychology*, *76*(4), 595–598.

Cendrillon, R., & Lovell, B. (2000). Real-time face recognition using eigenfaces. In *Visual communications and image processing* (Vol. 4067, pp. 269–276). International Society for Optics; Photonics.

Krumhuber, E. G., Küster, D., Namba, S., & Skora, L. (2020). Human and machine validation of 14 databases of dynamic facial expressions. *Behavior Research Methods*, 1–16. https://doi.org/https://doi.org/10.3758/s13428-020-01443-y

Lottridge, D., & Chignell, M. (2009). Emotional bandwidth: Information theory analysis of affective response ratings using a continuous slider. In *IFIP conference on human-computer interaction* (pp. 111–114). Springer.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71–86.