Supplementary Materials: EEG-Based Online Regulation of Difficulty in Simulated Flying

Ping-Keng Jao, Ricardo Chavarriaga, *Senior Member, IEEE*, and José del R. Millán, *Fellow, IEEE*

Abstract—This document provides the method in skill evaluation with more details. Besides, neural correlates, subject-wise results on behavioral results, offline validation, and closed-loop evaluation with more details are also provided.

Index Terms—EEG, Real-time decoding of difficulty, Closed-loop adaptation, (Simulated) Flying

S.I Details of Skill Evaluation

Figure S1 illustrates the process of personalizing difficulty levels which involves a collection of hit rates and sigmoid regression. The left side of Figure S1 shows the collection of data, and then, the lower part indicates that personalization is done by a sigmoid regression.

For collecting hit rates, the subject needed to navigate the drone through a pre-defined trajectory of eleven levels (radii), where each level has eleven waypoints and associate to one hit rate. Each hit rate corresponds to a blue data point on the right side of Figure S1. We then performed a sigmoid regression with the x-axis being the radius of waypoint and the y-axis being the hit rate. Thirteen radii were firstly sampled from the regression curve with 0%, 8.3%, ..., and 100% of hit rates. Three additional levels were included with the radii being 1.5, 2, and 2.5 times larger than the radius of 100% hit rate, in order to include some extremely easy levels. This resulted in defining sixteen levels in the recordings, and the sampled points are plotted as red dots in Figure S1.

The eleven radii used to personalize the sixteen radii in the offline recording were mostly 0.05, 0.1, 0.3, 0.5, 0.8, 1, 1.5, 2, 5, 8, and 10. For the four subjects who had lower skills and got lost during training, the radii were increased manually by the operator to make sure they can finish, where 0.1 is the lowest radius. While in the online sessions, the eleven levels were calculated from conducting another sigmoid regression based on the hit rates of the offline session. The offline session, as shown in Figure S2, has thirty-two trajectories, where each is associated with a difficulty level (waypoint size). As a result, the sigmoid regression for defining the eleven levels in the online sessions was computed from thirty-two data points, and the targeted eleven radii had the hit rates of 0%, 10%, ..., and 100%.

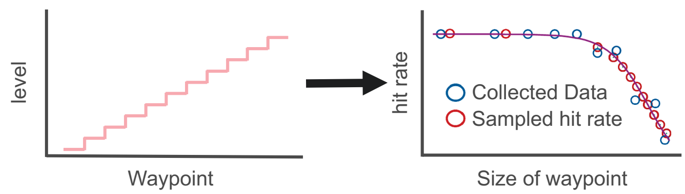


Figure S1: Personalizing objective difficulty is based on skill evaluation before the recording.

- *Ping-Keng Jao is with the École Polytechnique Fédérale de Lausanne, Geneva, Switzerland (email: ping-keng.jao@alumni.epfl.ch).*
- *Ricardo Chavarriaga is with the ZHAW Datalab, Zurich University of Applied Sciences, Winterhur, Switzerland, and also with the École Polytechnique Fédérale de Lausanne (email: ricardo.chavarriaga@zhaw.ch).*
- *José del R. Millán is with the Department of Electrical and Computer Engineering & the Department of Neurology, University of Texas at Austin and also with École Polytechnique Fédérale de Lausanne. (email: jose.millan@austin.utexas.edu).*

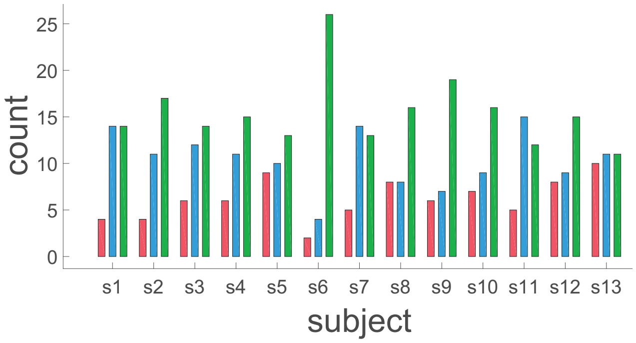
Manuscript received Seot. DD, 2020; revised Month DD, 2020. This work has been supported by the Swiss National Centres of Competence in Research (NCCR) Robotics. Digital Object Identifier 00.0000/????.0000.0000000

S.II Supplementary Results

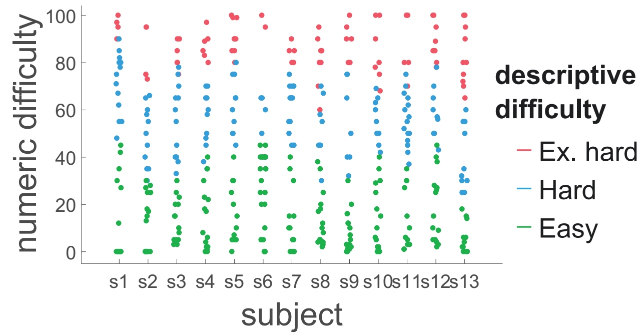
S.II.1 Offline Behavioral Result

Figure S2 summarizes the reported numerical and descriptive difficulty levels, and also hit rates of each subject, instead of grand average as in the main article. Figure S2(a) provides the number of trajectories being categorized as Easy, Hard, or Extremely Hard. Six subjects had an amount of Easy trajectories similar to the sum of the other two labels. Subject s6 had a much higher number of Easy trajectories because the personalization was conducted without sufficient training time during setup. Although the distributions across subjects were not necessarily the same, the protocol generally induced sufficient samples for validating under the targeted binary-classification framework. Figure S2(b) further scatters the numerical and descriptive labels, some intra-subject inconsistency can be observed in s8, s10, and s11, where a few numerical levels were described as Hard while another time as Extremely Hard. Roughly speaking, a level below 45 is considered as Easy while above 70 is considered as Extremely Hard.

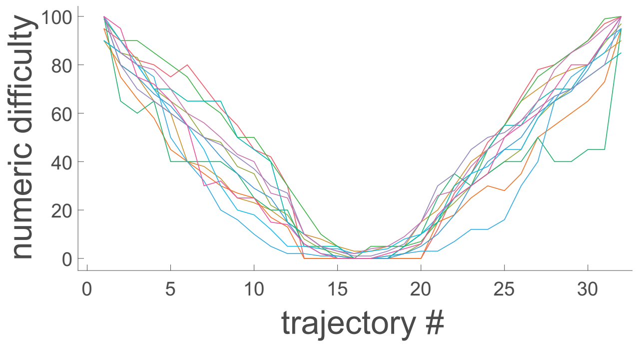
Figure S2(c) displays how the numeric levels evolved across trajectories, where one line stands for a subject, similar to the grand average of Figure 2 in the main article.



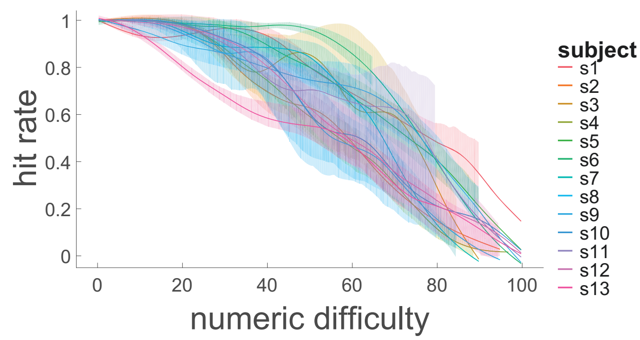
(a) Samples per descriptive label



(b) Numeric vs. descriptive labels



(c) Time profiles



(d) Hit rates vs. difficulty

Figure S2: Subject's reported difficulty labels and hit rates in the offline recording.

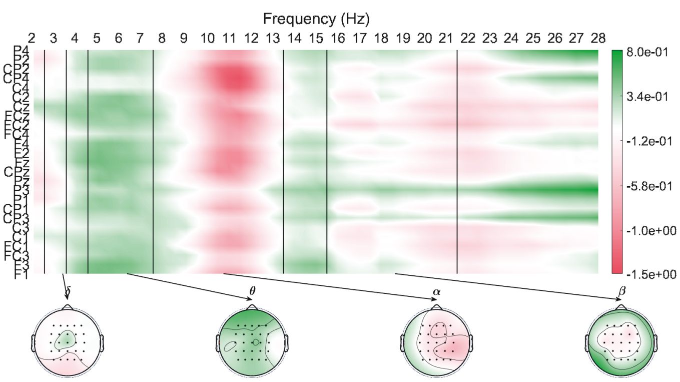


Figure S3: Differences between two grand averages of log-PSD: [(Ex. Hard & Hard) - Easy]. The grand average was first performed over windows, and then subjects. Red (Green) means that a lower value favors the Hard (Easy) condition. White means no difference.

Some smaller waypoint sizes were reported as easier than the previous larger one, but the trend in a larger scale shows that the difficulty levels changed as expected. Figure S2(d), on the other hand, illustrates the relationship between hit rates and reported numerical difficulty for each subject for the thirty-two trajectories. The data was spline-smoothed with the shaded areas as standard deviations [*S1*]. The hit rate generally decreases when the numeric difficulty level increases. The trend is similar to the right side of Figure 1 in the main article, although far from perfect and the x-axes are not the same.

S.II.2 Neural Correlates

Figure S3 shows the difference of the grand averaged log-PSD for the targeted classification, [(Ex. Hard, Hard), Easy], in the online sessions. The electrode-frequency plot illustrates the power for each frequency bin and electrode. White means there is no difference between the two conditions while the darker the red or green means the two conditions are easier to be distinguished. The topoplots below show the differences in four frequency bands: δ (3Hz), θ (4-7 Hz), α (8-13 Hz), and β (16-21 Hz). Visually, prominent features are the δ band around Cz, the θ band around Cz and frontal lobe, and the α band in the right hemisphere

S.II.3 Offline Accuracy Validation

The mean class-balanced accuracy in offline validation across subjects was 76.7% with the standard deviation being 5.1%. Figure S4 further depicts the per-class accuracy (green and red) and class-balanced accuracy (blue) at window level for each subject with her/his own best hyper-parameter. In the figure, the H(ard)+Ex.H(ard) class was grouped as one class according to the online interaction principle. The boxes provide the quartiles while the data points are test accuracies during the validation. It is easily observed that some test sets were often largely misclassified and appeared more frequently in the Easy class.

Figure S5 provides a clue to a possible explanation of this phenomenon. The red lines are hit rates and blue lines are the reported numerical difficulty levels. A cross stands for the accuracy of a tested trajectory and one color maps to a specific descriptive label. A trajectory may have multiple data points because of the validation strategy.

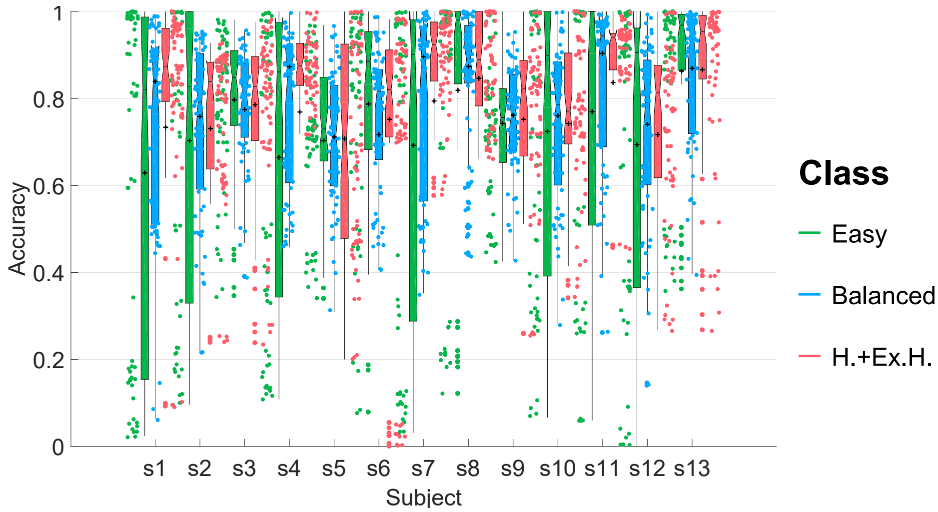


Figure S4: Accuracy at window level in offline validation.

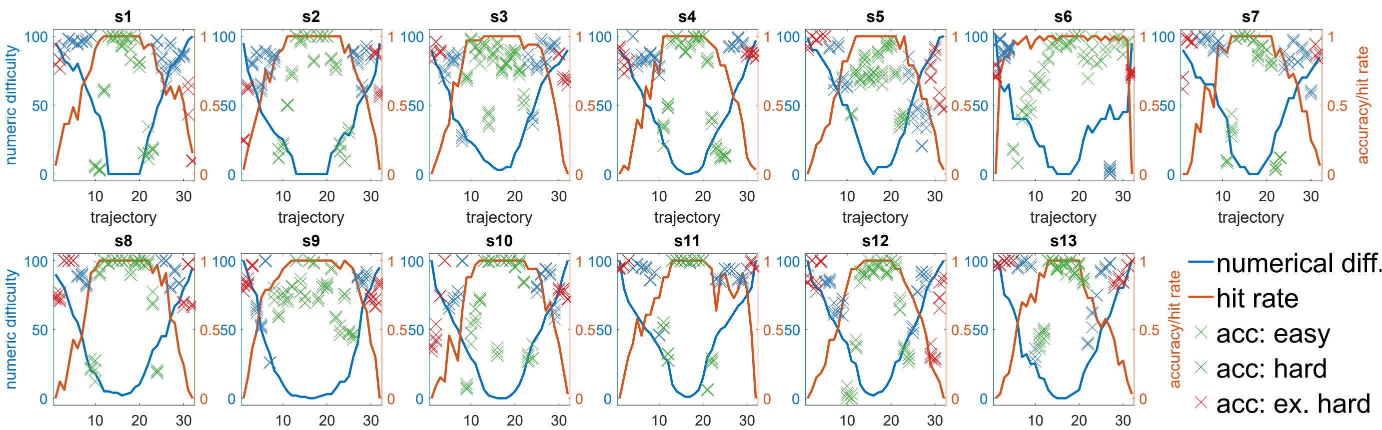


Figure S5: Accuracy at window level in offline validation per trajectory for each subject. The classification was Easy vs. Hard + Extremely Hard. Please noteice that the classification was done between Easy vs (Hard + Ex. Hard).

Taking s1 as an example, the accuracies at around 10th and 23rd trajectories suddenly drop. It is easily observed that these trajectories have a numeric difficulty level around the boundary between Easy (green) and Hard (blue). Similarly, s13 has the same situation at around 11th and 22nd trajectories. It has to be remembered that the classification was done between Easy v.s. (Hard + Ex. Hard). Therefore, the transitions between Hard (blue) and Ex. Hard (red) was mostly not affected as in s13. One exception can be found is around the 32rd trajectory of s1, where a potential explanation is that s1 was tired and disengaged, given that it was extremely hard and the long recording was about to finish.

The finding from the above two examples is that the data points with low test accuracies are mainly located between the transition trajectories, where the descriptive labels are switching from the Easy (green) to Hard (blue) with a few exceptions. Similar transitional trends can be identified from all the subjects. Therefore, we believe that the regression properly captured the near-ordinal trends of EEG correlates, but the responses of regression at transitional cases were too similar. As a result, the one-dimensional Linear Discriminant Analysis (LDA) was not able to perfectly separate them, but instead, yielded a threshold with the best class-balanced accuracy.

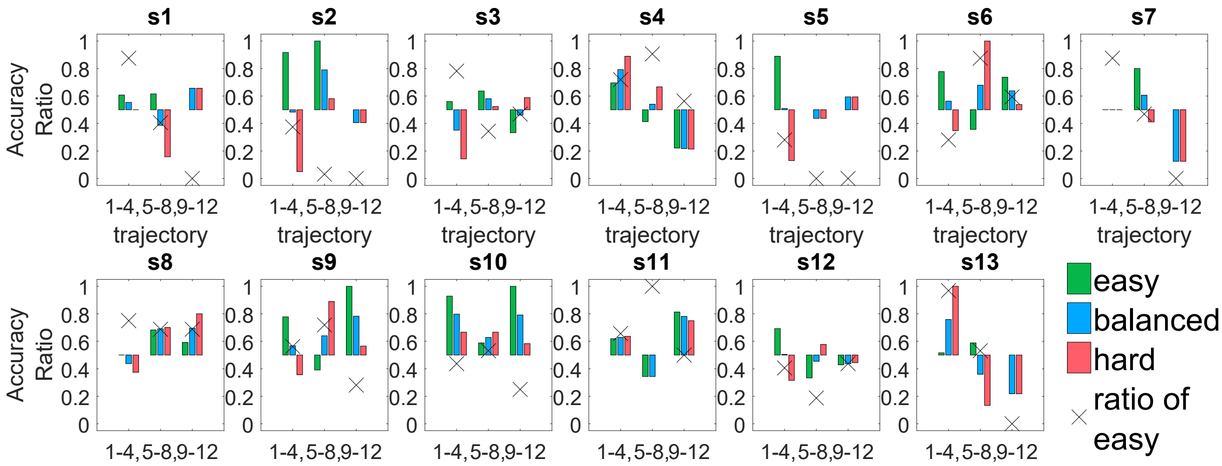
S.II.4 Online Accuracy Validation

Figure S6 depicts the accuracies during online validation by grouping four trajectories for each subject. Each group ensures the same bias term and the accuracy is based on 32 decision points. For both sessions, 50 out of 78 blue bars (˜64%) are above the chance level.

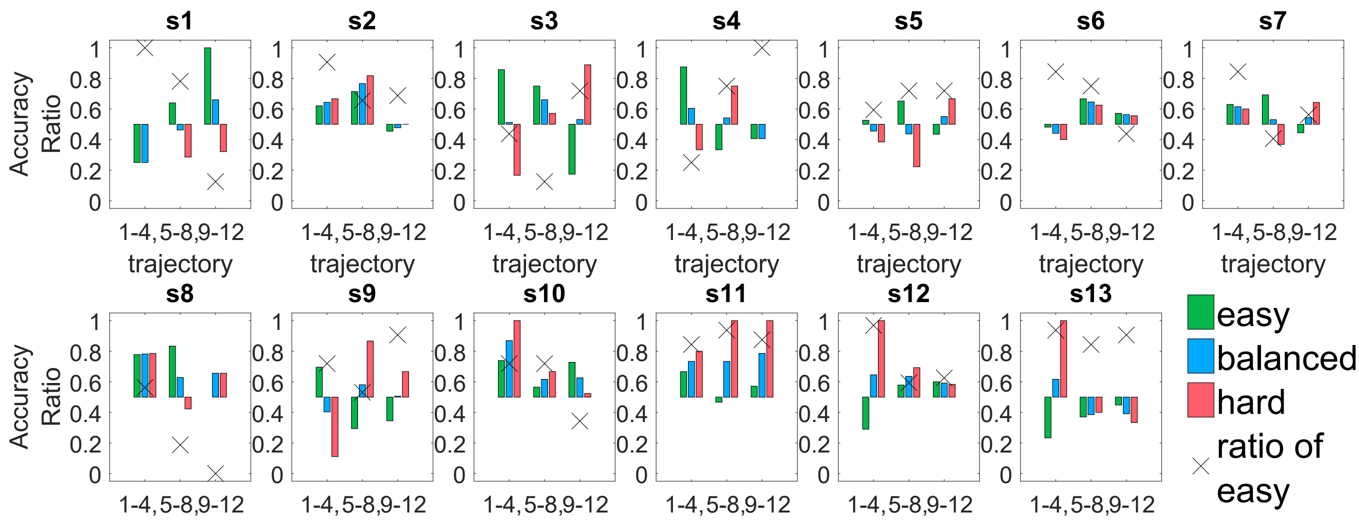
Figure S7 shows the used shift of bias term during online decoding, where zero means that the bias term of regression was the same as the best model learned from the offline session. The red bars show the recordings using the same shift for all the 12 trajectories, in total, 9 out 26 kept the same shifts, where s10 had zero shifts across both recordings.

S.II.5 Online – Final Levels

Figure S8 illustrates the curves representing the final level of each trajectory for each subject and session. The red curves stand for the Manual condition while the blue ones are for the EEG condition. As defined in Section 5.3, the correlation, MD (mean difference), and MAD (mean of absolute difference) are provided for each subject as texts. The



(a) Session 2



(b) Session 3

Figure S6: Online decoding accuracy in EEG condition at decision-point level, grouped by four trajectories to ensure the same shift of the bias term in regression.

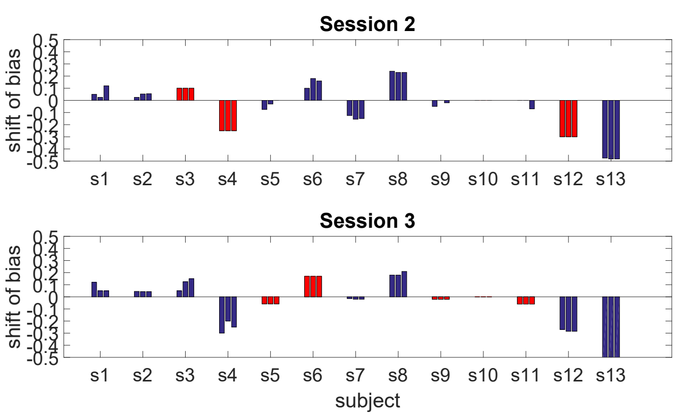
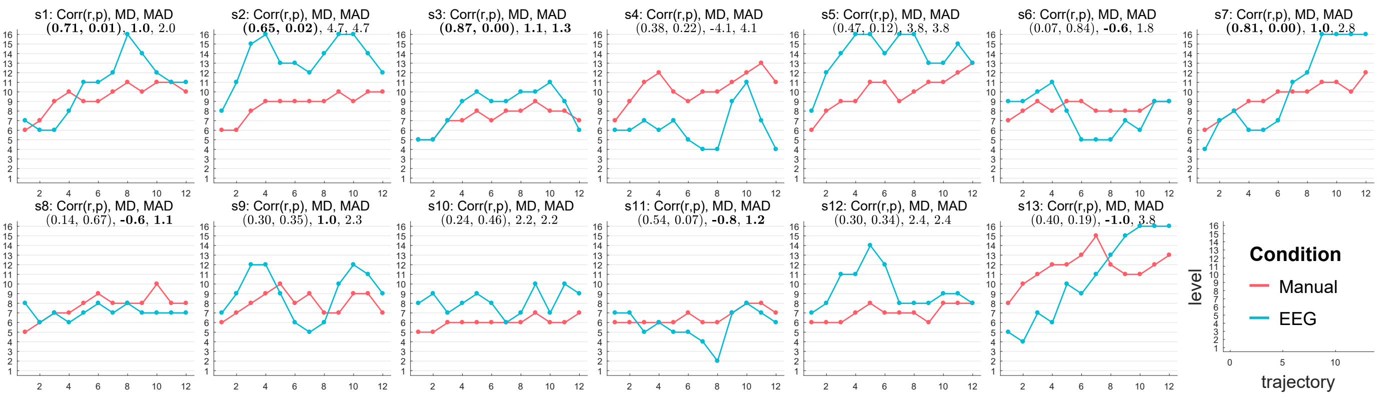


Figure S7: Shift of the bias term in online sessions, where red indicates the same shift in a session.

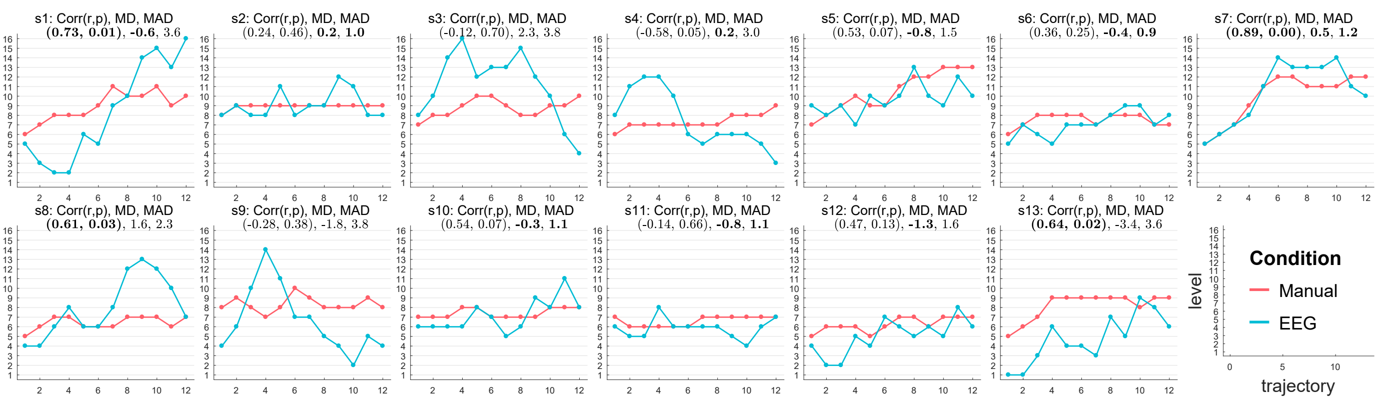
boldfaced values are correlation coefficients larger than 0.5 with statistical significance, and also for $|\mathrm{MD}|$ (MAD) which are below 1.5 level (level[2]). The proposed indices behaved as explained in Section 5.3. For example, subject s11 has very good values in MD and MAD for both sessions but the r is rather low. Referring to the curves in both sessions, s11 has the patterns of EEG conditions being similar to Manual ones. However, a few overshoots are over-emphasized by the correlation coefficients. In total, five and six subjects in session 2 and session 3, respectively, have two indices indicating high similarities.

References

[S1] P. Morel, "Gramm: Grammar of graphics plotting in Matlab." *J. Open Source Software*, vol. 3, no. 23, pp. 568–571, 2018.



(a) Session 2



(b) Session 3

Figure S8: Final levels across trajectories.