Instructions to analysts EEGManyPipelines project

Dear Analyst,

Welcome to the EEGManyPipelines project!

Please carefully read the instructions below. The instructions will inform you on 1) the dataset to analyse, 2) the hypotheses to answer, 3) the type of analyses to perform, and 4) what and how to report your results and outcome of analyses at the end of the analysis phase (April 30th 2022).

Please complete the prior expectations questionnaire linked in the email you received after reading these instructions and the data documentation. Note that every individual analyst (i.e. not just one member of a team) has to fill in this questionnaire. Afterwards, you are ready to download the data and start analyzing!

A submission portal will be available towards the end of the analysis phase, including questionnaires and options for submitting codes and pre-processed data. This submission portal is not yet open, but it will be opened in March 2022. More details forthcoming.

If you have any questions, please contact us via committee@eegmanypipelines.org. Please also consult our FAQ page (https://www.eegmanypipelines.org/) if you have questions or encounter problems. We will keep the FAQ page updated throughout the entire project.

Best wishes,
The EEGManyPipelines steering committee

Description data set **2**

Hypotheses **2**

How to analyze the data **3**

Instructions for submission of results, code, data **4**

A) Instructions for the results questionnaire 4
B) Instructions for free-text results report 4
C) Instructions for analysis questionnaire 5
D) Instructions for how to submit your analysis scripts 5
E) Instructions for how to submit your pre-processed data 7

Description data set

The dataset includes raw EEG data from a study on memory for visual scenes from 33 subjects recorded with a 70 channel EEG system. The images showed either man-made environments or natural environments. Some images were repeated throughout the experiment and the subjects had to report whether the images were “old” (had appeared previously) or “new” (had not appeared previously).

A detailed description of the dataset, EEG channel layout, triggers, etc. can be found in the download repository in the “documentation” folder in the document “EMP_dataset_documentation.pdf”. Please read the data documentation carefully before you begin the data analysis.

The data can be downloaded using this link:

https://uni-muenster.sciebo.de/s/zg2U8xVNZbRHqzg

Hypotheses

The objective of your data analysis is to test the following hypotheses:

1. There is an effect of *scene category* (i.e., a difference between images showing man-made vs. natural environments) on the amplitude of the N1 component, i.e. the first major negative EEG voltage deflection.

2. There are effects of *image novelty* (i.e., between images shown for the first time/new vs. repeated/old images) within the time-range from 300–500 ms ...
    a. ... on EEG voltage at fronto-central channels.
    b. ... on theta power at fronto-central channels.
    c. ... on alpha power at posterior channels.

3. There are effects of *successful recognition* of old images (i.e., a difference between old images correctly recognized as old [hits] vs. old images incorrectly judged as new [misses]) ...
    a. ... on EEG voltage at any channels, at any time.
    b. ... on spectral power, at any frequencies, at any channels, at any time.

4. There are effects of *subsequent memory* (i.e., a difference between images that will be successfully remembered vs. forgotten on a subsequent repetition) ...
    a. ... on EEG voltage at any channels, at any time.
    b. ... on spectral power, at any frequencies, at any channels, at any time.

Please note:

- All timing-related specifications above refer to the time of image onsets, i.e. the time points indicated by trigger markers in the EEG data files.
- The hypotheses refer to the experimental conditions (scene category, old, behavior, subsequent memory) that are coded in the EEG trigger markers, as explained in the dataset documentation.

- "EEG voltage" refers to the conventional time-domain signal. "Theta power" and "alpha power" refer to the response within the canonically defined alpha and theta bands in the EEG literature. “Spectral power” refers to the result of a (time-)frequency transform (e.g. FFT, wavelets, multitapers, etc.).
- Some of these hypotheses are intentionally vague, thus there might be several different, plausible ways in which the data could be analyzed and the hypotheses tested. We would like you to use similar analysis procedures that you would normally use in your own studies.
- The terms and concepts in the hypotheses above (e.g. N1 component, spectral power, theta power, etc.) should be familiar to EEG researchers. We will not provide further specification about the data and hypotheses beyond what is written above, in the data documentation document, and in the online FAQ (https://www.eegmanypipelines.org/faq.php).

How to analyze the data

Our primary aim in this project is to assess the variety of EEG analysis pipelines “in the wild”. To this end, we want to set as few guidelines or restrictions as possible to achieve the goal. It is up to you how you want to process the data and test the hypotheses.

- The most important guideline is: your analysis pipeline (including preprocessing and hypothesis testing) should correspond as **closely** as possible to the way you would usually analyze **your own EEG data**! Please approach the hypotheses as if they were your own research project and take all the steps you deem necessary to test the hypotheses similar to how you normally would do.

- It is required to test these specific hypotheses. In the results report, we will ask you whether you confirmed each respective hypothesis or not. There will be other opportunities to indicate more “shades of grey” about your results; but in principle, we ask you to make a binary “yes”/ “no” decision.

- The hypotheses concern fairly standard EEG analysis, which EEG researchers should be familiar with. You are not expected to analyze anything beyond what is required to test the hypotheses or to use any methods that are more “advanced” than what you are using currently for analyzing data from your own experiments. We are also not interested in additional “exploratory analyses” that go beyond testing the hypotheses stated above.

- The data are provided as raw data. The data have been only minimally preprocessed (please refer to the dataset documentation). This implies that we have not cleaned the data; they still include the full range of artefacts (e.g. bad channels, eye-blinks, cardiac artefacts, etc.) that can be realistically expected. You are free to discard data of those subjects, channels, or time segments that you deem “low quality”. Again, your definition of “low quality” and the procedures used for dealing with low-quality data should correspond as closely as possible to your usual procedures. Also note that in your report, we will ask you for all your procedures (including code) and for the exact indices of discarded subjects, channels, and trials (see below).

Instructions for submission of results, code, data

At the end of the analysis period, each team is expected to submit the results and materials listed below. To help you manage your objectives, here's an overview of what you will be required to submit:

A. A questionnaire collecting your result for each hypothesis and how confident you are about these results.
B. A short report detailing your results (free-text, similar to a results section in a scientific paper).
C. A questionnaire collecting detailed information about the analysis procedures adopted by your team.
D. Your analysis scripts (if you used scripts to process the data).
E. Your pre-processed trial-by-trial data.

The submission portal with the questionnaires and platform where you must submit your analysis scripts and pre-processed data will open in March 2022. More details will follow when we get closer to the submission phase.

Below you will find further descriptions and guidelines on how to prepare for the submission, so you know what to prepare and have ready for the data submission phase.

A) Instructions for the results questionnaire

In this questionnaire you will report whether you confirmed each hypothesis or not by providing a “yes” or “no” answer. You have to make a binary choice because this project asks you to test specific hypotheses. In addition, you will rate your confidence in your result for each respective hypothesis and estimate the percentage of teams that will have confirmed the respective hypothesis. This questionnaire will be similar to the prior expectations questionnaire you will fill in at the beginning of the analysis phase (until November 1st, 2021).

B) Instructions for free-text results report

We ask you to report your results for each hypothesis using a free-text result report. The report should be concise and formatted similar to the results section of a research scientific journal article (around one paragraph per hypothesis). The report should, at the very least, include the decisive statistical parameters (e.g. test-statistic, p-value, or other decision criterion) that guided your decisions on whether the hypotheses were confirmed or not. Please be prepared to report numerical values with a higher level of numerical precision than you would usually do (e.g. 7–8 digits after the comma or a numerically exact fraction). Depending on how you would generally write up results, you can include additional information, e.g. when and where condition differences arise, how strong they are, etc. You can also have a brief discussion about your findings (e.g. qualifiers, follow-up and control analyses,...) if applicable.

C) Instructions for analysis questionnaire

In this questionnaire, you will be asked a series of questions about how you analyzed the data. Given that these reports will form the core data of the *EEGManyPipelines* project, we will ask you to provide as many details as possible. To facilitate filling out this document, we designed most of the questions as multiple choice. However, in case you believe the alternatives we provided are insufficient, there will be free-text fields where you can provide more information. If you perform a certain step at several time points during your analysis pipeline, there will be ways to accommodate that.

We recommend that you take note and document all your analysis steps while working on your analysis, so that this information will be easy to retrieve later when filling out this questionnaire.

D) Instructions for how to submit your analysis scripts

We will also ask you to submit the analysis scripts used for each step of your pre-processing pipeline. The goal is for the steering committee to be able to provide the identical results by executing exactly the same analysis scripts.

We would like you to submit the scripts in **one of the following formats** depending on the analysis software you used:

- **I**f you used any self-written **scripts**, please provide all those scripts (for software-specific details, see the list below). If these scripts in turn used functions from publicly available software packages or toolboxes, do not submit these toolboxes! It is enough that you tell us which toolboxes and versions you used. Please make sure that your code clearly indicates **the order** in which your scripts were called (e.g. in a seperate text file or by numbering the scripts, (e.g. *01-filtering*, *02-ica*, etc.). Given that we may have to work with your code for our own analyses, we would appreciate it if your scripts were coded as transparently as possible. Please pay special attention to documenting how you did the time-frequency decomposition.
- If you use (mainly) the **Graphical User Interface** (GUI) of the analysis software, you should provide a text file (e.g. docx, pdf) in which you describe your steps. You can use both free text and screenshots. Such a text file should contain the values chosen for each entry field of a given preprocessing step. Explicitly report which parameters you decided to leave empty (‘default values’), see Figure 1 and 2. It can be useful to include descriptive pictures/ screenshots with the details of the graphical windows, as illustrated in Figure 1 and 2.
  Also, many software packages allow you to export and include an *ordered list of functions* (with all the input settings). Such files can easily be generated by most EEG analysis software through a *“history” command* or *script generator*. Please see below for software specific guidelines, and refer to the documentation for the software you are using, on how to export history scripts.

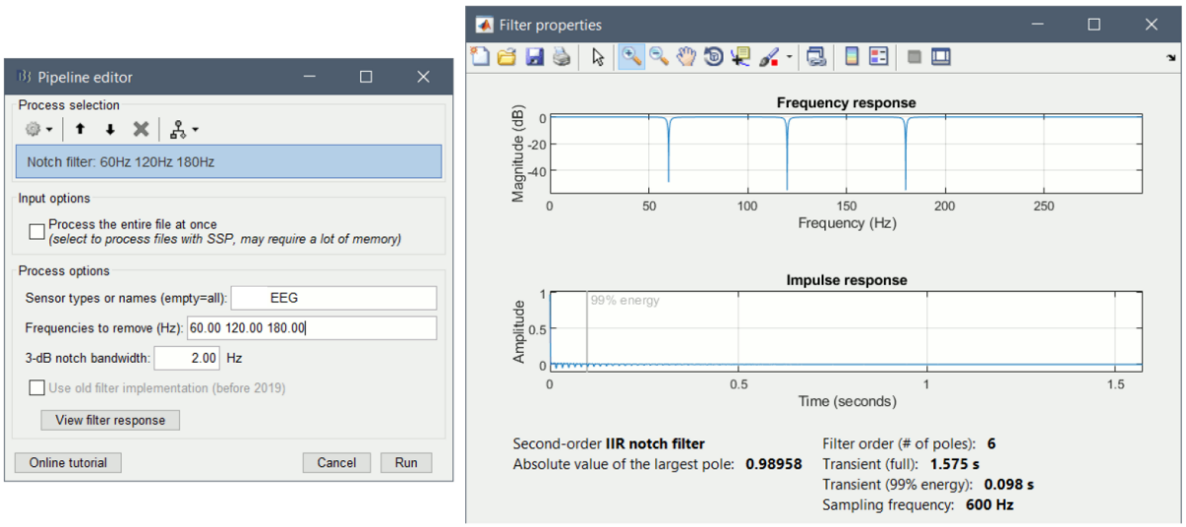

Figure 1. An example for the "notch filter" editor as implemented in *Brainstorm* software,

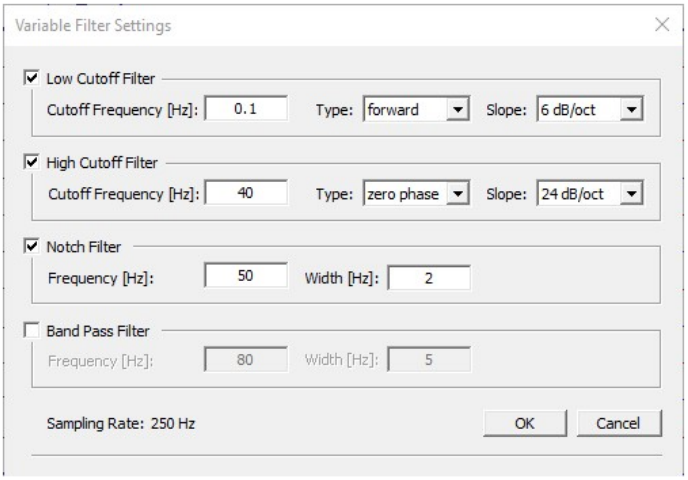

Figure 2. An example for the "Variable Filter Settings" window as implemented in *BESA* software.

Below are toolbox specific guidelines for code submission for some of the most commonly used EEG analysis software:

- **Brainstorm**: please use the script generator in the pipeline editor to extract the sequence of functions applied to the data. Alternatively you can provide a clear description with pictures (screenshots) of each step parameters applied into a .doc or .txt file (as in the example above, see Figure 1).
- **BrainVision Analyzer:** for each operation (i.e. pre-processing step), please copy and paste the respective log files into a .txt file. You can retrieve this log information by right-clicking on the operation and selecting "Operation infos". Note that the operation infos will not be saved and written to the log file until after the processing is completed.
- **EEGLAB (GUI)**: EEGLAB, LIMO, and ERPLAB use command history functions that automatically store any mouse clicks in the GUI as equivalent script functions. If your analysis was exclusively GUI based, please check that the command history was

recorded accurately and completely. Please share the complete command history with us (.set files, .txt file).

- **EEGLAB (scripts)**: please provide all your custom scripts (.m files) and specify whether other plugins or toolboxes are required to run your scripts. If you use a combination of automatic code-based analysis and manual GUI-based analysis, make sure to provide your code as well as the automatically recorded command history in the data files.
- **Fieldtrip**: Please provide all your analysis scripts (.m files), including any custom scripts.
- **MNE-python**: please provide all your analysis scripts (.py and/or .ipynb), including any custom scripts.
- **SPM-EEG:** once you have analyzed the dataset with the GUI or batch system, use the "spm_eeg_history" function to automatically generate the script with the list of all the processing steps and input arguments.
- **Custom Matlab** code: provide all Matlab custom scripts used.
- **Custom Python** code: provide all Python custom scripts used.
- **Custom R** code: provide all R custom scripts used.
- **If none of the above applies to your analyses procedure, please contact us via email to committee@eegmanypipelines.org before starting the analysis.**

E) Instructions for how to submit your pre-processed data

Once you have finished the entire analysis, we will ask you to submit the **trial-by-trial pre-processed data for each subject** at the end of the data analysis phase. As for the script instruction, the format for sharing the data should allow us to extract the relevant variables/ information for cross-teams comparison. Please submit all text responses and comments in **English**. In most cases, use the native data format of the software you use; if in doubt, use .txt or .csv files.

We require you to create a **folder for each subject** (e.g. *subj01*, *subj02,* …). Each folder should contain the data and data reports for one specific subject. Do this for all subjects in the dataset, even if you excluded subjects before the final analysis. Each subject folder must contain:

1. The single-subject **time-domain data** at the end of the pre-processing steps (i.e. after any data cleaning and preprocessing steps, but **before any averaging** and inference). The data will in most cases be a 3D structure with **channels x samples x trials**. If your analysis is based on continuous data (i.e. without segmentation in trials or "'epoching'), please submit your data in a 2D format with channels x samples. All pre-processed data of one subject should be stored within one single file. The data should also include the trigger value for each trial (see data documentation). This information is stored within the data structure or data file as a default for many analysis software. You therefore just have to make sure the data is in the right format and submit the data in the native format used by your analysis software (e.g. .fif, .set., .mat, etc.).

   If the trigger values are not contained in the output file of your analysis software, you have to generate a text file (.txt or .csv) with one column reporting the corresponding

trigger value for each epoch. This list with trigger values should have the same length as the number of trials in your preprocessed data.

2. A text file (.txt or .csv) with the **number/label of channels** that you identified as "bad channels" in the data analysis (e.g. channels you removed and/or eventually interpolated) **for each subject**. The channels not reported as "bad channels" are assumed to be included in the following pre-processing and further analysis.

3. A text file (.txt or .csv) with the **indices of the trials rejected** during the pre-processing **for each subject** (e.g. Epochs rejected: 2, 5, 9, etc.). Important note: if you rejected epochs in two or more separate steps, you should report in the first row the indices of the epochs rejected during the first step (by specifying the analysis step), in the second row the epochs rejected in a second step, and so on.

   Example:
   (line 1): Trial indices rejected at step 1 (before ICA): 2, 5, 9, ….
   (line 2): Trial indices rejected at step 2 (after ICA): 2, 23, 48, ….
   Please report the trial indices *relative to all trials* (i.e., not separated by conditions) and retain consistent trial indices across preprocessing steps. For example, if you reject trial 1 at an early step, the index of trial 2 should remain 2 (and not get updated to 1).

   If data were analyzed in a continuous format without segmentation or "epoching" in trials, create a plain text file (.txt or .csv) with two columns: the first with the sample index corresponding to the beginning of the removed segment, and the second column with the sample index corresponding to the end of the removed segment.

   Example:
   (line 1): 5362, 9678
   (line 2): 25641, 35414
   ...

4. In case you performed an **Independent Component Analysis**, please provide a full-text description (e.g., .txt, .docx, .pdf) **for each subject** with the following information:
   a. The total number of ICs **obtained** from ICA decomposition
   b. The overall number of ICs that you **selected** for removal
   c. A breakdown with the number of components you identified as pertaining to a given class of non-brain signal. Report only the classes you decided to identify and remove. Please use the following suggested classes: a) **eye blink**; b) **eye-movement** (e.g. lateral eye-movement); c) **muscular artifact;** d) **channel noise**; e) **line noise**; f) **heart**; f) **other** (please specify in a few words).

   An example of a report should be: "*For subject 1, our ICA decomposition yields 42 components. From those, we rejected a total of 6 components, with 2 being eyeblinks; 3 muscle artifacts, and 1 being heart-related*".