Is EEG better left alone for decoding?

Roman Kessler [1,*] colleagues[2],

1 Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

*** rkesslerx@gmail.com**

Abstract

TODO

Introduction

TODO

Materials and Methods

Datasets

For the analysis, I used the awesome and openly available `erpcore` dataset [8], downloaded from OSF. 40 participants underwent six different well-studied ERP experiments to isolate the following components: ERN, LRP, MMN, N170, N2pc, N400, and P3 [8]. The ERN and LRP were examined using the same paradigm. EEG was recorded using 30 scalp electrodes (Fig. S1), along with three electrooculogram (EOG) channels, two positioned lateral to the outer canthus of each eye and one below the right eye [8]. No participants or sessions were excluded due to artifacts, as different forking paths would result into different participants being excluded by each forking path. However, the original authors' processing of the dataset used exclusions based on artifacts or low performance, resulting in a maximum of 6 out of 40 participants being excluded for a single paradigm [8]. For the rest of the experimental design and paradigms, I refer to Kappenman et al. (2021) [8], and I will only address changes to their processing or aspects that seem important for the purposes of this manuscript. In the remainder of this article, I will refer to these datasets by their component names (e.g., MMN, N170, etc., as shown in Table 1).

Data pre-preprocessing

Each dataset was processed using `MNE` ([5], v. 1.5.1) in Python. Triggers from the raw signal were pruned to retain only relevant events (Table 1. ERN and LRP trials were analyzed relative to button presses, while all others were analyzed relative to stimulus onset. Since decoding was performed on all available scalp electrodes rather than on individual electrodes, some experiments were repurposed. For example, in LRP I divided the conditions into left and right button presses. Following the textbook intuition of LRP, a subtraction of contra- and ipsilateral electrodes would have been performed, which would have prevented the possibility of decoding from all scalp

channels. Similarly, for N2pc, I divided the conditions into target left and target right for decoding (Table 1).

Stimulus event times were shifted forward by 26ms to account for the delay of the LCD monitor. The data was downsampled from 1024 to 256 Hz. The raw data channels remained in single-ended mode, so no reference was yet applied. The signal from the two horizontal EOG channels was subtracted to form a single horizontal EOG channel (HEOG). Similarly, the signal from Fp2 was subtracted from the vertical EOG channel (positioned below the right eye) to form a single vertical EOG channel (VEOG). These data were then used for multiverse processing.

Table 1. Datasets. Properties of the `erpcore` data sets [8]. For each data set, the conditions and channels used to compute and visualize the evoked responses are shown, where the second condition is subtracted from the first condition. The conditions are also repurposed for decoding in some experiments for convenience.

dataset	paradigm	conditions (evoked)	conditions (decoding)	channels (evoked)
ERN	Eriksen flander task [4]	incorrect, correct	incorrect, correct	FCz
LRP	Eriksen flander task [4]	contralateral, ipsilateral	response left, response right	C3/C4
MMN	passive auditory oddball	deviants, frequents	deviants, frequents	FCz
N170	face perception	faces, cars	faces, cars	PO8
N2pc	visual search	contralateral, ipsilateral	target left, target right	PO7/PO8
N400	word pair judgment	unrelated, related	unrelated, related	CPz
P3	active visual oddball	deviants, frequents	deviants, frequents	Pz

Multiverse preprocessing

All preprocessing steps are restricted to functions that are either included or closely related to the `MNE` package, or at least written in Python. Several processing steps have been identified that vary across the EEG literature. However, we closely follow the processing steps also investigated by [1, 2, 12]. This includes high pass filtering (hpf), low pass filtering (lpf), eye movement correction using independent component analysis (ICA), muscle artifact correction using ICA, re-referencing, detrending, baseline correction, and artifact correction using `autoreject` [6, 7]. See Table 2 for an overview of the different analysis options. A total of 1152 forking path resulted from the systematic variation of all processing steps.

Some processing steps could have been extended with other methods. For example, there are various possible methods for eye movement correction. However, I decided not to include too many options for all processing steps in order to keep the computation time within reasonable limits. Also, I assume that several eye movement correction methods would lead to very similar signals. Furthermore, the order of the different processing steps was not varied, as this would have led to a combinatorial explosion that would have required much higher computational resources. In my opinion, varying the order of many processing steps (such as rereferencing and filtering) effectively results in similar signals.

XXX evaluate more single examples

Only processing steps that did not change the number of epochs in the final dataset were performed. In particular, artifact-contaminated trials were corrected rather than rejected. Steps such as eye-movement correction only included methods that do not reject but rather correct artifact-contaminated segments. Similarly, the hyperparameters in the `autoreject` package were set to values that interpolated rather than rejected noisy trials. First, this ensured a balanced number of epochs per condition for decoding in experiments with a similar number of trials per condition. This included

Table 2. Variation of analysis steps. 1152 forking paths evolved by the systematic variation of each analysis step.

step	variations
re-referencing	average, Cz, P9/P10
high pass filter (hpf)	None, .1Hz, .5Hz
low pass filter (lpf)	None, 6Hz, 15Hz, 45Hz
eye movement correction	None, ICA
muscle artifact correction	None, ICA
detrending	offset, linear
baseline correction	200ms, 400ms
autoreject [6,7]	False, True

N170, N2pc, N400, and LRP (nearly similar number of trials). Second, it allowed for a more intuitive interpretation of the decoding scores, in a way that model performance was not driven by different numbers of included trials per forking path. Of course, there are also good reasons for a different approach at this point, which can be investigated in the future, such as comparing decoder performance between trial rejection and trial correction as another interesting split in the multiverse, but this will not be discussed here. Similarly, no trials were excluded due to behavioral performance, such as incorrect responses, contrary to what was done in [8] for all experiments except ERN, or due to high reaction times.

XXX TODO: when the EEG Many labs paper is out, compare with the processing steps there!

Referencing. The channels of the raw data were re-referenced to either a single channel (i.e., Cz), the average of two channels near the mastoids (P9, P10), or the average of all channels (Tab. 2). P9 and P10 were chosen according to [8]. According to their previous studies, the average of P9 and P10 provided cleaner signals than the commonly used mastoids [8].

Filtering. The default `filter` function of `MNE` was used on all channels of raw time series, applying a linear FIR filter with high and low pass cutoffs defined by the respective forking paths (Tab. 2).

Eye-Movement Correction. In forking paths with eye movement correction, `ICA` was performed using the corresponding `MNE` function. First, the raw signal was copied and the copy was high-pass filtered at $1Hz$. Then an ICA was performed with the `picard` method, setting the maximum number of iterations to 500 and estimating 20 components. The `find_bads_eog` function with default parameters was used to correlate each component with the artificially generated EOG channels HEOG and VEOG (see above). Components were classified as EOG signals by passing a threshold based on adaptive z-scoring. The ICA solution was then applied to the original (unfiltered) data, effectively subtracting the problematic components.

TODO: couldn't find original references

Muscle artifact correction. The same procedure as for eye movement correction was performed, except that `find_bads_muscle` is used to determine muscle artifacts [3,11].

Epoching. The previous processing steps were performed on the continuous signal. Epoching was performed for a time window of $1.2s$ for each experiment (Tab.S1). The interval included the same interval as [8] used for each experiment, but added $.2s$ at the end for ERN and LRP, or added $.2s$ at the beginning for all other experiments (Tab.S1). The extension was made to allow for a variation (i.e., expansion) of the baseline period in the multiverse, and to keep the signal lengths equal between paradigms. The exact parameters vary for each forking path (Tab. 2).

Detrending. Detrending was performed prior to baseline correction. Both operations are carried out on each individual channel of each epoch. In the first approach, a constant is subtracted from each epoch and channel (intercept only or "offset"). Alternatively, linear detrending is performed, where a linear function with intercept and slope is fitted to each epoch and channel using least squares, and the signal is detrended by continuing only with the corresponding residuals.

Baseline correction. A baseline correction is then applied by calculating the mean of the specified baseline period time interval for each channel and epoch. This mean is then subtracted from the remainder of the epoch for that channel. The baseline period varied between forking paths (Tab. S1).

Artifact correction using `autoreject` For further artifact detection and interpolation, the `autoreject` package (v. 0.4.2) was used on the epochs of some forking paths [6, 7]. In short, the package automatically detects and rejects or interpolates bad sensors and trials based on peak-to-peak values. The rejection thresholds are determined by cross-validation. In the current study, I set the `autoreject` hyperparameters to values such that all bad sensors are interpolated rather than rejected. This ensured that the resulting number of trials for each forking path remained the same. One could also investigate the sweet spot between rejection and decoding accuracy, but that is left for future studies.

Evoked responses

For each experiment, exemplary grand average evoked responses were computed as a proof of principle to ensure that the data processing worked as expected. Therefore, for each experiment, I chose a forking path that is loosely related to the processing in the original paper that presented the datasets ([8]). That is, for all experiments except N170, re-referencing was done on mastoids (average of P9 and P10), with an HPF of .1, and no no LPF. ICA was used for both eye and muscle artifact correction, as well as `autoreject`, since the original study performed multiple (including manual) artifact correction steps. Detrending was done using the offset, and the baseline period was $200ms$. For N170, re-referencing was done using a mean reference, similar to [8]. The remaining steps were the same for all experiments.

For LRP and N2pc, the channels for which evoked responses were computed were combined to obtain signals from contralateral and ipsilateral to the target (N2pc) or response (LRP) (Table 1). This was done only for visualization of evoked responses for comparison with [8], not for decoding analysis.

Decoding models

TODO Einführung in typische decoding approaches, trial wise, time resolved, feature extraction then decoding, frequency wise...

Trial-wise decoding using `EEGNet`

The first decoding model applied was `EEGNet` (version 4) [9] as implemented in the `braindecode` toolbox (version .8) [5, 10]. For each forking path, the preprocessed trials were each standardized using exponential mean and variance (i.e., `exponential_moving_standardize` with default parameters). Because some of the experiments had an unbalanced number of trials per condition, class weights were computed separately for each condition, experiment, and participant. The MMN, P3, and ERN experiments had highly unbalanced classes, while the LRP experiment was slightly unbalanced. Next, the order of the trials was shuffled and a stratified 5-way cross-validation split was defined. The model was defined with default parameters, a batch size of 16, and the maximum number of epochs was set to 200. Scoring was done using balanced accuracy for the 2-class case (eq. 1), which effectively reduces to the usual accuracy measure for balanced datasets. For each cross-validation split, the model was fitted to $\frac{4}{5}$ of the data and evaluated on the remaining $\frac{1}{5}$ of the data. The resulting validation scores were averaged. This procedure was repeated for each forking path, experiment, and participant.

$$balanced\ accuracy = \frac{sensitivity + specificity}{2} \tag{1}$$

Time-resolved decoding using `Logistic Regression`

EEG literature is full of time-resolved decoding analyses (XXX REF). This approach typically uses the electrode activations of each time point separately to train respective classifiers (XXX REF). Sometimes, the single trials are averaged into several pseudo evoked potentials for each condition, such that in e.g., a 3-fold cross-validation procedure, 2 evokeds for each class are in the training set, and the remaining evoked of each class is put in the test set (XXX Liu, Bae, 3. ausm JC). To get some precise accuracy estimation despite the lack of enough test data, several independent iterations of shuffling are conducted so that the composition of underlying epochs for each pseudo-evoked differ in each run. Model performances are then averaged across shuffling iterations and cross-validation steps.

In the present analysis, the epochs are used without averaging, as done e.g., in (XXX). A simple logistic regression estimator us used

Analysis of forking path performance

Results

Evoked responses

Evoked responses were calculated for each experiment using an exemplary bifurcation path (Fig. 1). The components closely follow the results of the original study (see [8] and its Fig. 2).

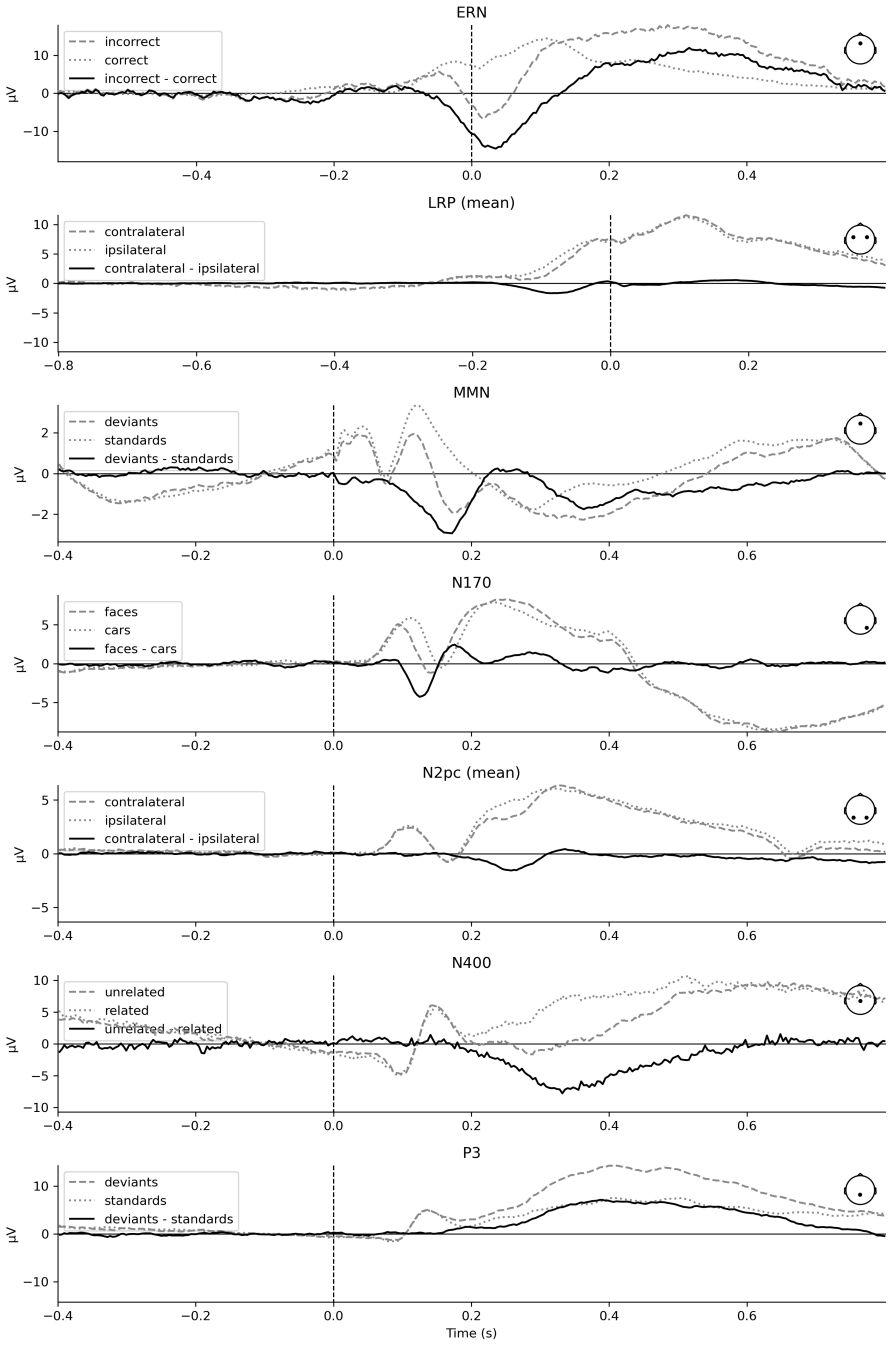

Figure 1. Grand average evoked responses for each experiment. The dashed and dotted lines represent the average responses to the respective stimulus categories. The solid line represents the difference between the respective categories, containing the components of interest. The time series derive either from one single channel, or from two channels, in which case the mean was calculated. The channel positions are represented by a small graphical legend in each subplot.

Discussion

Subsection heading.

TODO

Supporting Information

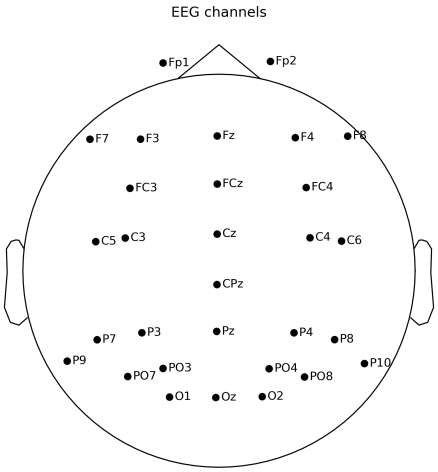

Figure S1. Applied EEG montage. 30 scalp electrodes were recorded alongside with three EOG electrodes (not shown).

Table S1. Epoching. Time-locking event, epoch lengths and baseline intervals. The total duration of the epochs is kept similar across experiments. The baseline interval varied per forking path, either to a window of 200ms or 400ms.

dataset	time-locking event	epoch interval [s]	baseline interval [s]	
			200ms	400ms
ERN	response	$(-.6; .6)$	$(-.4; -.2)$	$(-.6; -.2)$
LRP	response	$(-.8; .4)$	$(-.8; -.6)$	$(-.6; -.2)$
MMN	stimulus	$(-.4; .8)$	$(-.2; .0)$	$(-.4; .0)$
N170	stimulus	$(-.4; .8)$	$(-.2; .0)$	$(-.4; .0)$
N2pc	stimulus	$(-.4; .8)$	$(-.2; .0)$	$(-.4; .0)$
N400	stimulus	$(-.4; .8)$	$(-.2; .0)$	$(-.4; .0)$
P3	stimulus	$(-.4; .8)$	$(-.2; .0)$	$(-.4; .0)$

References

1. P. E. Clayson, S. A. Baldwin, H. A. Rocha, and M. J. Larson. The data-processing multiverse of event-related potentials (ERPs): A roadmap for the optimization and standardization of ERP processing and reduction pipelines. *NeuroImage*, 245:118712, Dec. 2021.

2. A. Delorme. EEG is better left alone. *Scientific Reports*, 13(1):2372, Feb. 2023.

3. D. Dharmaprani, H. K. Nguyen, T. W. Lewis, D. DeLosAngeles, J. O. Willoughby, and K. J. Pope. A comparison of independent component analysis algorithms and measures to discriminate between EEG and artifact components. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 825–828. IEEE, 2016.

4. B. A. Eriksen and C. W. Eriksen. Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, 16(1):143–149, Jan. 1974.

5. A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hämäläinen. MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7(7 DEC):1–13, 2013.

6. M. Jas, D. Engemann, F. Raimondo, Y. Bekhti, and A. Gramfort. Automated rejection and repair of bad trials in MEG/EEG. In *6th International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, Trento, Italy, June 2016.

7. M. Jas, D. A. Engemann, Y. Bekhti, F. Raimondo, and A. Gramfort. Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, 159:417–429, Oct. 2017.

8. E. S. Kappenman, J. L. Farrens, W. Zhang, A. X. Stewart, and S. J. Luck. ERP CORE: An open resource for human event-related potential research. *NeuroImage*, 225:117465, Jan. 2021.

9. V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, Oct. 2018.

10. R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, Aug. 2017.

11. E. M. Whitham, K. J. Pope, S. P. Fitzgibbon, T. Lewis, C. R. Clark, S. Loveless, M. Broberg, A. Wallace, D. DeLosAngeles, P. Lillie, A. Hardy, R. Fronsko, A. Pulbrook, and J. O. Willoughby. Scalp electrical recording during paralysis: Quantitative evidence that EEG frequencies above 20Hz are contaminated by EMG. *Clinical Neurophysiology*, 118(8):1877–1888, Aug. 2007.

12. A. Šoškić, S. J. Styles, E. S. Kappenman, and V. Kovic. Garden of forking paths in ERP research – effects of varying pre-processing and analysis steps in an N400 experiment. preprint, PsyArXiv, Mar. 2022.