Bring a map when exploring the ERP data processing multiverse: A commentary on Clayson et al. 2021

Daniel Feuerriegel*, Stefan Bode

Melbourne School of Psychological Sciences, The University of Melbourne, Melbourne, Australia

**A R T I C L E   I N F O**

**Keywords:**
Electroencephalography
EEG
Event-related potentials
ERPs
Error-related negativity
Error positivity
Multiverse analyses

**A B S T R A C T**

Clayson et al. (2021) describe an innovative multiverse analysis to evaluate effects of data processing choices on event-related potential (ERP) measures. Based on their results, they provide data processing recommendations for studies measuring the error-related negativity and error positivity components. We argue that, although their data-driven approach is useful for identifying how data processing choices influence ERP results, it is not sufficient for devising optimal data processing pipelines. As an example, we focus on the inappropriate use of pre-response ERP baselines in their analyses, which leads to biased error positivity amplitude measures. Results of multiverse analyses should be supplemented with further investigation into why differences in ERP results occur across data processing choices before devising general recommendations.

Electroencephalographic (EEG) data must often be processed before it can be analysed. Data processing procedures, by their nature, alter EEG data and influence our ability to accurately measure signals of interest, such as event-related potential (ERP) components. Clayson et al. (2021) propose a way to systematically assess the impacts of different processing choices on ERP measures using multiverse analyses. In their paper, the same large dataset was processed in multiple ways using different processing choices or parameter settings. Metrics of ERP measures were then compared across data processing choices. Processing pipelines diverged with respect to the high- and low-pass filters used, the method used to correct for eye movement artefacts, the reference electrode(s), the ERP baseline window, the electrode(s) used for ERP component amplitude measurement, and the type of measurement used (e.g., peak or mean amplitude measures).

They demonstrate the utility of this approach by focusing on the error-related negativity (ERN) and error positivity (Pe) components, which differ in amplitude between trials with correct and erroneous decisions in many paradigms (e.g., Falkenstein et al., 1991; Gehring et al., 1993). Specifically, they aimed to identify processing steps that i) minimised across-trial amplitude variability and ii.) maximised differences in ERP measures across correct responses and errors. Based on their results, they recommend optimal data processing choices for measuring the ERN and Pe components in future work. They promote their multiverse analysis approach as useful for deriving optimal processing pipelines for other components in relation to different research questions.

In this commentary, we highlight important caveats of multiverse analyses that should be considered when making general recommendations for EEG data processing choices. We believe that the approach in Clayson et al. (2021) and related work using multiverse analyses (e.g., Sandre et al., 2020; Klawohn et al., 2020; Šoškić et al., 2022) is useful for identifying how ERP results differ across different processing pipelines. This, in turn, can help us understand why effects do (or do not) replicate across studies that processed their data in different ways. However, we advise caution when basing data processing and ERP measurement choices solely on the results of multiverse analyses such as those in Clayson et al. (2021). These optimisation-based approaches favour processing choices that produce measurement bias, and in some cases can even lead to spurious results. To demonstrate this point, we focus on the choices of ERP baseline windows used for measuring the Pe in their multiverse analyses.

1. Accounting for bias in multiverse analyses

A subset of the multiverse analyses in Clayson et al. (2021) were designed to identify processing choices that produce maximal ERP amplitude differences between correct and error trials. One major shortcoming, however, is that these analyses do not detect or estimate measurement biases that can occur due to certain data processing choices. Here, we discuss measurement bias in terms of observed ERP component amplitude differences between correct and error trials. Biases can arise because the procedure is applied to real datasets where we do not

---

* Corresponding author at: Melbourne School of Psychological Sciences. Redmond Barry Building, The University of Melbourne, 3010, Australia.  
E-mail address: dfeuerriegel@unimelb.edu.au (D. Feuerriegel).

https://doi.org/10.1016/j.neuroimage.2022.119443.

Received 3 April 2022; Received in revised form 30 May 2022; Accepted 1 July 2022
Available online 2 July 2022.
1053-8119/© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
know the true magnitude of ERN or Pe amplitude differences across correct and error trials. Clayson and colleagues identify one such source of bias in their results: that peak amplitude (as opposed to mean amplitude) measures maximised ERN differences between correct and error trials. They rightly state that this result likely occurred because there were fewer epochs included per participant for errors compared to correct responses, which leads to inflated measures of peak amplitude in error trials (Thomas et al., 2004). The authors acknowledge that multivariate analyses may sometimes lead to settings that are suboptimal or inappropriate, and that care should be taken when interpreting these results.

However, these sources of bias are not always easy to identify post hoc. Here, we describe another source of bias relevant to measuring Pe amplitudes: the choice of the ERP baseline window. Clayson et al. (2021) included a range of pre-response baselines in their multivariate analyses, whereby the average amplitude over the specified time window was subtracted from each trial before averaging ERPs. They found that a baseline window of −200 to 0 ms preceding each keypress response maximised [error − correct] differences in Pe amplitudes. They recommended that this baseline window should be used in future work. The same set of baseline windows was also used in the multivariate analyses of Sandre et al. (2020) and Klawohn et al. (2020) who focused on the ERN component.

Baseline correction is based on the assumption that there are no systematic differences in ERPs across conditions during the baseline period. In perceptual decision tasks, such as those used to elicit the ERN and Pe, there are often differences in ERP amplitudes across trials with correct responses and errors during the pre-response baseline time window (e.g., Feuerriegel et al., 2021a), which have their origin in the cognitive processes leading up to response initiation. These effects are typically captured by analyses of the centro-parietal positivity (CPP) component (O’Connell et al., 2012) that rises to a peak immediately before each keypress response at the same parietal electrodes that are used to measure the Pe. When there are systematic differences in ERPs during the pre-response baseline window, those differences will be projected with opposite polarity to the post-response period of the epoch (Luck, 2014), including the time window at which the Pe is typically measured (e.g., 200–400 ms post response). For example, when pre-response ERPs are more positive-going for correct responses as compared to errors, a pre-response baseline will produce a bias toward relatively more negative-going Pe amplitudes for correct responses, and more positive-going amplitudes for errors. This, in turn, will artificially inflate measures of Pe [error − correct] difference scores.

More generally, the assumption of equivalent pre-response amplitudes across correct and error trials is implausible for many commonly used perceptual decision tasks. Response time (RT) distributions differ across correct and error trials in most experiments (Ratcliff et al., 2016; Ulrich et al., 2015), meaning that the pre-response baseline window will overlap with different portions of stimulus-locked ERP waveforms in each condition (discussed in Sandre et al., 2020). This can produce measurement biases in either direction, depending on the influence of the stimulus-locked ERPs during the pre-response baseline windows.

Importantly, this issue can often be avoided by using baselines that precede the onset of the stimulus in each trial (pre-stimulus baselines). This method of baseline-correction is performed by first deriving epochs time-locked to the stimulus, subtracting a pre-stimulus baseline, and then deriving response-locked epochs from the resulting data. This approach relies on the assumption that ERPs do not systematically differ during the pre-stimulus baseline period and should only be used when this assumption is plausible (see Alday, 2019, for discussion and an alternative, regression-based method). Issues associated with pre-response baselines are prevalent across the error-processing literature and such baselines are conventionally used in this area of research. We encourage researchers who are building their data processing pipelines to be critical of this convention.

In fact, the use of pre-response baselines can, in some cases, produce spurious differences across conditions. In other datasets, it can artefactually inflate (or deflate) ERP difference measures. To illustrate each of these effects we have plotted ERPs time-locked to keypress responses for trials with correct responses and errors from four separate datasets (Fig. 1). These were selected because they are openly available and include ERPs that can be corrected using both pre-stimulus and pre-response baselines.

The first dataset (Feuerriegel et al., 2021a) is from an experiment using a difficult perceptual discrimination task that did not present arrays of incongruent stimuli. When a pre-stimulus baseline is used (Fig. 1A), there are clear differences in ERPs during the pre-response time window, but not in the subsequent Pe measurement window. After applying a pre-response baseline, the differences in pre-response ERPs are propagated to the post-response period, creating a spurious difference in Pe amplitudes between correct and error trials (Fig. 1B). The second dataset (Feuerriegel et al., 2021b) used a modified Flanker task that presented orientated gratings. Data are from trials with incongruent target/flanker arrays. There is a clear Pe effect for both pre-stimulus and pre-response baselines (Figs. 1C, 1D). However, the pre-response baseline artefactually inflates the size of this effect. The third dataset (Figs. 1E-F) is from a typical Flanker task taken from the ERP CORE database (Kappenman et al., 2021). In this dataset, Pe effects were inflated when using pre-response baselines. Finally, we present a subset of 40 (randomly selected) datasets from Bode and Stahl (2014) using a more typical Flanker task. ERPs have been current source density-transformed, but are still relevant for illustrating our point. Here, pre-response ERP amplitudes are more positive-going for errors (Fig. 1G), artefactually deflating effect sizes when using a pre-response baseline (Fig. 1H).

We formally tested for differences in time window mean amplitudes between correct and error conditions during pre-response baseline windows, using ERPs corrected using a −100 to 0 ms pre-stimulus baseline. In these analyses, more positive-going amplitudes for correct conditions during the pre-response baseline window will result in an inflation of the [error − correct] difference during the subsequent Pe measurement window if a pre-response baseline is applied. The results (shown in Table 1) were in line with the effects indicated by Fig. 1. When measuring amplitudes between −100 to 0 ms relative to the response, more positive-going amplitudes were observed for correct trials in Feuerriegel et al. (2021a, 2021b) and Kappenman et al. (2021). More negative-going amplitudes were instead observed for the dataset of Bode and Stahl (2014). When a slightly longer −200 to 0 ms measurement window was used (as recommended by Clayson et al., 2021), the same patterns were observed, except that the difference in the Feuerriegel et al. (2021b) dataset was not statistically significant.

Here, we note that real [error − correct] Pe amplitude differences have been observed when using pre-stimulus baselines (e.g., Murphy et al., 2015) and are not simply an artefact of pre-response baseline correction. We also note that this issue is not specific to the −200 to 0 ms baseline recommended by Clayson et al. (2021), as different biases may also be apparent during earlier time windows (e.g., the −500 to −300 ms window included in their analyses). During these earlier time windows, there are amplitude differences which reflect different build-up rates of the CPP (e.g., O’Connell et al., 2012; Steinemann et al., 2018; Feuerriegel et al., 2021b). For example, conditions with faster RTs show steeper CPP build-up rates (i.e., steeper ERP slopes) that start rising closer to the time of the response. By contrast, slower RTs produce a more gradual rising amplitude slope that begins further back in time. This can produce more positive-going ERPs in conditions with slower RTs (for example trials with slow errors that occur in Flanker tasks, see Ulrich et al., 2015) around −500 to −300 ms prior to the response (e.g., Figure 4B in Feuerriegel et al., 2021b; Figure 4G in Steinemann et al., 2018). This will, in turn, produce measurement bias when the pre-response baseline overlaps with the time window of these amplitude differences, and when there are RT differences across correct
trials and errors, as is often the case (Ratcliff et al., 2016; Ulrich et al., 2015).

This issue is also relevant to another goal stated in Clayson et al. (2021) and also Sandre et al. (2020), which is to use multiverse analyses to standardise processing pipelines for studies using psychometric approaches. In the examples described above, the use of pre-response baselines essentially conflates effects during the pre-response baseline with those during the Pe measurement window of interest. For example, if some individuals show larger [correct – error] differences in pre-response ERPs, this can be mistaken for larger effects on Pe amplitudes, even if there are no true differences during the Pe measurement window. This, in turn, could artefactually inflate (or obfuscate) individual differences for a component of interest, and influence estimates of the between-person variability of ERP measures.

Bias can also arise when using other optimisation criteria, such as minimising across-trial amplitude variability. For example, analyses us-
ing this metric in Clayson et al. (2021) also favoured the −200 to 0 ms pre-response baseline choice for the P3 component.

2. Recommendations for using multiverse analyses

Here, we have highlighted an instance where multiverse analyses lead to recommendations of processing choices that produce biased estimates of PE amplitude effects. The direction and extent of this bias can differ across experiments (and individuals), and in some instances, the use of pre-response baselines can even produce spurious ERP differences across experimental conditions (e.g., Feuerriegel et al., 2021a). The optimisation-based approach in Clayson et al. (2021) was not designed to identify (or correct for) such sources of bias and will tend to favour processing steps that inflate ERP effect magnitudes. Importantly, such biases are not always easy to identify by an experimenter post hoc.

Based on these observations, we recommend multiverse analyses as a valuable tool for assessing whether results are robust across sets of data processing choices (Sandre et al., 2020; Klawohn et al., 2020; Clayson et al., 2021; Šoškič et al., 2022). If there are substantial differences across processing choices, this warrants systematic investigation of why such differences occurred, either through simulation studies or further analyses of relevant features of the data. For example, effects of pre-response baselines could be assessed by deriving single-trial ERPs corrected using pre-stimulus baselines, and then testing for systematic differences in ERP amplitudes during the pre-response baseline window in each trial. Specification curve analysis (Simonsohn et al., 2020) may also be useful for identifying specific data processing choices that warrant further scrutiny. Systematically investigating why results differ may uncover features of the data that do not conform to the assumptions underlying certain data processing or analysis options. Documenting these features could help us build more appropriate pipelines in future work.

However, we do not recommend the approaches in Clayson et al. (2021) or related work (Sandre et al., 2020; Klawohn et al., 2020; Šoškič et al., 2022) as the primary method for identifying optimal data processing pipelines, as the optimisation procedure can favour processing choices that bias ERP component measures. As we have shown here, common knowledge of ERP data features is not always sufficient to identify such sources of bias post hoc. When using real data, we do not know the true underlying effect magnitudes, either within individuals or at the group level. In some cases we do not know whether the specific experimental manipulation actually produces the hypothesised effect. When there is no real effect to be found, there is the risk of experimenter bias toward selecting processing choices that produce a set of hypothesised effects.

In summary, we believe that the data-driven multiverse analysis proposed by Clayson et al. (2021) is a valuable tool for identifying how differences in data processing choices lead to differences in ERP results. However, rather than optimising across-condition differences or across-trial variability measures, we believe the guiding principle for developing EEG processing pipelines should be that of identifying and minimising measurement bias. This can help to ensure that multiverse analyses do not lead us astray, even in cases where there are strong conventions in the field that favour certain processing choices, such as using pre-response baselines.

Data and code availability statement

EEG datasets used to create the figure in this commentary are freely available at osf.io/gazx2/, osf.io/eucqf/, osf.io/thsgg/ and osf.io/bndjg/. Code used to reproduce the plots in Fig. 1, as well as averaged ERP data, is available from osf.io/guwnm/.

Acknowledgements

Writing of this commentary was supported by an Australian Research Council Discovery Early Career Researcher Award to D.F. (ARC DE220101508). Funding sources had no role in the development of the ideas presented in this commentary. We thank Patrick Cooper for his insightful feedback on an earlier draft.

References

Alday, P.M., 2019. Extended GLM model can replace baseline correction while lifting its limits. Psychophysiology 56 (12), e13451.
Bode, S., Stahl, J., 2014. Predicting errors from patterns of event-related potentials preceding an overt response. Biol. Psychol. 103, 357–369.
Clayson, P.E., Baldwin, S.A., Rocha, H.A., Larson, M.J., 2021. 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.
Falkenstein, M., Hohnsbein, J., Hoormann, J., Blanke, L., 1991. Effects of crossmodal divided attention on late ERP components II. Error processing in choice reaction tasks. Electroencephalogr. Clin. Neurophysiol. 78 (6), 447–455.
Feuerriegel, D.C., Murphy, M., Konski, A., Sun, J., Mepani, V., Hester, R., Bode, S., 2021a. Electrophysiological correlates of confidence differ across correct and erroneous perceptual decisions. bioRxiv doi:10.1101/2021.11.22.469610.
Feuerriegel, D., Jiwa, M., Turner, W.F., Andrejevic, M., Hester, R., Bode, S., 2021. Tracking dynamic adjustments to decision making and performance monitoring processes in conflict tasks. Neuroimage 238, 118265.
Gehring, W.J., Goos, B., Coles, M.G., Meyer, D.E., Donchin, E., 1993. A neural system for error detection and compensation. Psychol. Sci. 4 (6), 385–390.
Kappenman, E.S., Farrens, J.J., Zhang, W., Stewart, A.X., Luck, S.J., 2021. ERP CORE: An open resource for human event-related potential research. Neuroimage 225, 117465.
Klawohn, J., Meyer, A., Weinberg, A., Hajicak, G., 2020. Methodological choices in event-related potential (ERP) research and their impact on internal consistency reliability and individual differences: An examination of the error-related negativity (ERN) and anxiety. J. Abnorm. Psychol. 129 (1), 29.
Luck, S.J., 2014. An Introduction to the Event-Related Potential Technique. MIT Press, Cambridge, MA.
Murphy, P.R., Robertson, L.H., Harry, S., O’Connell, R.G., 2015. Neural evidence accumulation persists after choice to inform metacognitive judgments. Elife 4, e11946.
O’Connell, R.G., Do clave, P.M., Kelly, S.P., 2012. A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. Nat. Neurosci. 15 (12), 1729–1735.
Ratcliff, R., Smith, P.L., Brown, S.D., McKoon, G., 2016. Diffusion decision model: Current issues and history. Trends Cogn. Sci. 20 (4), 266–281.
Sandre, A., Banica, I., Riesel, A., Flake, J., Klawohn, J., & Weinberg, A. (2020). Comparing the effects of different methodological decisions on the error-related negativity and its association with behaviour and gender. Int. J. Psychophysiol., 156, 18–39

Table 1

Comparisons of Mean Amplitudes Between Correct and Error Trials During Windows Used for Pre-Response Baselines.

| Dataset | t    | df | Correct - Error Mean Difference | p    |
|---------|------|----|---------------------------------|------|
| −100 to 0 ms Pre-Response Baseline Window |
| Feuerriegel et al. (2021a) | 7.08 | 28 | 1.35 μV | < 0.001 |
| Feuerriegel et al. (2021b) | 3.41 | 27 | 0.72 μV | 0.002 |
| Kappenman et al. (2021) | 5.08 | 35 | 2.49 μV | < 0.001 |
| Bode & Stahl (2014) | −3.24 | 39 | −2.23 μV/m² | 0.002 |
| −200 to 0 ms Pre-Response Baseline Window |
| Feuerriegel et al. (2021a) | 6.57 | 28 | 1.30 μV | < 0.001 |
| Feuerriegel et al. (2021b) | 1.42 | 27 | 0.34 μV | 0.168 |
| Kappenman et al. (2021) | 3.84 | 35 | 1.62 μV | < 0.001 |
| Bode & Stahl (2014) | −2.15 | 39 | −1.17 μV/m² | 0.038 |
Simonsohn, U., Simmons, J.P., Nelson, L.D., 2020. Specification curve analysis. Nat. Hum. Behav. 4 (11), 1208–1214.
Šoškić, A., Styles, S.J., Kappenman, E.S., Kovic, V., 2022. Garden of forking paths in ERP research—Effects of varying pre-processing and analysis steps in an N400 experiment. PsyArxiv doi:10.31234/osf.io/8rjah.
Steinemann, N.A., O’Connell, R.G., Kelly, S.P, 2018. Decisions are expedited through multiple neural adjustments spanning the sensorimotor hierarchy. Nat. Commun. 9 (1), 1–13.
Thomas, D.G., Grice, J.W., Najm-Briscoe, R.G., Williams Miller, J., 2004. The influence of unequal numbers of trials on comparisons of average event-related potentials. Dev. Neuropsychol. 26 (3), 753–774.
Ulrich, R., Schriber, H., Leuthold, H., Birngruber, T., 2015. Automatic and controlled stimulus processing in conflict tasks: Superimposed diffusion processes and delta functions. Cognit. Psychol. 78, 148-174.