Received: 6 September 2018 | Revised: 7 June 2019 | Accepted: 9 June 2019

DOI: 10.1111/psyp.13451

ORIGINAL ARTICLE

PSYCHOPHYSIOLOGY SPR WILEY

How much baseline correction do we need in ERP research? Extended GLM model can replace baseline correction while lifting its limits

Phillip M. Alday

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Correspondence
Phillip M. Alday, Max Planck Institute for Psycholinguistics, Postbus 310, 6500 AH Nijmegen, The Netherlands.
Email: phillip.alday@mpi.nl

Abstract

Baseline correction plays an important role in past and current methodological debates in ERP research (e.g., the Tanner vs. Maess debate in the *Journal of Neuroscience Methods*), serving as a potential alternative to strong high-pass filtering. However, the very assumptions that underlie traditional baseline also undermine it, implying a reduction in the signal-to-noise ratio. In other words, traditional baseline correction is statistically unnecessary and even undesirable. Including the baseline interval as a predictor in a GLM-based statistical approach allows the data to determine how much baseline correction is needed, including both full traditional and no baseline correction as special cases. This reduces the amount of variance in the residual error term and thus has the potential to increase statistical power.

KEYWORDS

analysis/statistical methods, EEG, ERPs, oscillation/time frequency analyses

1 | INTRODUCTION

Baseline correction belongs to one of the standard procedures in ERP research (cf. Luck, 2005) yet comes with two inherent difficulties: the choice of baseline interval and the assumption that there are no systematic differences between conditions in the baseline interval. Often discussed in conjunction with high-pass filtering, baseline correction is argued to be an artifact-free way to compensate for signal drifts in electrophysiological recordings (cf. the recent debate started in the *Journal of Neuroscience Methods*: Maess, Schröger, & Widmann, 2016a, 2016b; Tanner, Morgan-Short, & Luck, 2015; Tanner, Norton, Morgan-Short, & Luck, 2016; Widmann, Schröger, & Maess, 2015). In the following, we will demonstrate that, regardless of the choice of baseline interval or high-pass filter setting, traditional baseline correction is never an optimal procedure with modern statistical methods. In short, the correct way to address potential bias introduced by signal drifts is by including the baseline period in the statistical analysis.

2 | THE GENERAL LINEAR MODEL IN ERP RESEARCH

At the heart of all common analyses in ERP research, whether repeated measures analysis of variance (ANOVA) or various forms of explicit regression, is the general linear model (GLM):

$$y = \sum_{i \in \text{covariates}} \beta_i x_i + \varepsilon$$

$$\varepsilon \sim N\left(0, \sigma^2\right) \tag{1}$$

where y represents a column vector of observed EEG data (usually averaged over a given time window and in ANOVA-based approaches, averaged over trials), x_i are column vectors of various predictors and covariates, β_i represents the (statistically determined) weights of the x_i, and ε represents the error term (i.e., residuals, which are assumed to be normally

distributed). In its usual form, the error term is assumed to be homogenous, (i.e., having the same variance across the entire model and thus independent of any particular observation—the homoskedacity assumption). In the case of baseline-corrected statistics analyses, we can decompose the *y* column into

$$y = y_{\text{window}} - y_{\text{baseline}} \tag{2}$$

(Note that it does not matter whether the baseline is subtracted from the entire epoch before averaging within a given time window or afterward. This is because the baseline correction for a given epoch is a constant and the average of the difference is the same as the difference to the average.)

This means that we can re-express our GLM as

$$y_{\text{window}} - y_{\text{baseline}} = \sum_{i \in \text{covariates}} \beta_i x_i + \varepsilon, \tag{3}$$

which we can further rewrite as

$$y_{\text{window}} = \sum_{i \in \text{covariates}} \beta_i x_i + y_{\text{baseline}} + \varepsilon \tag{4}$$

To highlight the fact that the baseline correction is now on the "predictors" side of the equation, we change its name from y_{baseline} to x_{baseline}:

$$y_{\text{window}} = \sum_{i \in \text{covariates}} \beta_i x_i + x_{\text{baseline}} + \varepsilon \tag{5}$$

We note that this is just a special case of a linear model with the baseline correction as a predictor, with the special case that $\beta_{\text{baseline}} = 1$ (and no baseline correction is exactly the case that $\beta_{\text{baseline}} = 0$). This already suggests a more general way forward: we make the baseline interval a proper predictor and allow the model to determine the weight empirically. Nonetheless, let us consider the usual assumptions of classical baseline correction.

3 | THE UNDERLYING ASSUMPTIONS OF TRADITIONAL BASELINE CORRECTION MAKE IT IRRELEVANT

For traditional baseline correction to be valid, we assume that experimental conditions (whether traditional discrete, factorial conditions or "continuous conditions" in more naturalistic and less parametric designs) do not differ systematically in the electrophysiological activity in their respective baseline intervals. If they were to differ systematically in their baseline interval, then traditional baseline correction would move effects from the baseline window into window of interest (cf. e.g., Luck, 2005). Component overlap between trials presents a particular set of problems for this assumption (cf. Luck, 2005), although component overlap within trials is also problematic, and several methods have been proposed to address this issue (Smith & Kutas, 2014a, 2014b). In the following, we will ignore this particular problem for simplicity and without loss of generality.

As we have assumed no systematic differences in the baseline interval between conditions, we can think of the vector of baseline values as noise: $x_{\text{baseline}} \sim N\left(\mu_{\text{baseline}}, \sigma^2_{\text{baseline}}\right)$, which we assume to be normally distributed without loss of generality. In this case, our linear model simplifies to:

$$y_{\text{window}} = \sum_{i \in \text{covariates}} \beta_i x_i + \bar{x}_{\text{baseline}} + \varepsilon' \tag{6}$$

where

$$\varepsilon' \sim N\left(0, \sigma^2 + \sigma^2_{\text{baseline}}\right) \tag{7}$$

In other words, under these assumptions, traditional baseline correction increases the variance of the error term, (i.e., increases the noise) without otherwise impacting the inferential engine beyond introducing a shared offset $\bar{x}_{\text{baseline}}$, which will typically be expressed as a change in the intercept term. However, we have made a small yet potentially misleading equivalency, namely, that "no systematic differences in the electrophysiological activity in the baseline interval" is the same as "no systematic differences in the baseline interval." Other physical and environmental differences may lead to conditions differing systematically in their baseline interval. In the case that they differ only in their mean, then the previous observation holds, although the offset introduced by the baseline is now conditional on the experimental condition (i.e., there is now an interaction term with condition). If, however, the variance of the baseline interval differs, then we no longer meet the assumption of homoskedacity, as the resulting error term ε' is not homogenous across conditions.

We note at this point that the mathematics of traditional baseline correction—subtracting out a reference signal—are the same as the mathematics for rereferencing. It is no surprise then that baseline correction suffers the same pitfalls as a bad reference, such as biasing apparent topographies and introducing noise (cf. Maess, Schröger, & Widmann, 2016a, 2016b; Urbach & Kutas, 2006). However, unlike rereferencing, where each channel is shifted by the same time-dependent value and thus the relative values remain the same even if the individual values change (see Figures 1 and 4 in Lau, Stroud, Plesch, & Phillips, 2006, for an example), baseline correction shifts each channel by a different time-independent signal and can change the observed topography. As such, even more so than the choice of reference, the choice of baseline influences the inferences that can be made about observed effects (see Section 5 below for further discussion on the choice of baseline window).

Finally, this result also holds for analyses of spectral power (ERSP) under the usual normalization procedure. In particular, the usual normalization of dividing the target window by the baseline window and then taking the logarithm of the quotient (i.e., converting to decibels) yields the same statistical model:

$$\log \frac{y_{\text{window}}}{x_{\text{baseline}}} = \log y_{\text{window}} - \log x_{\text{baseline}}. \quad (8)$$

In the following, we will omit further explicit mention of time-frequency analyses, but we note that all results and suggestions apply equally to ERP and ERSP.

4 | EXPLICIT REGRESSION ON SINGLE-TRIAL DATA AS AN OPTIMAL SOLUTION

Returning to explicit regression without the baseline window included in the error term, we can consider the simple case of one experimental predictor[1]:

$$y_{\text{window}} = \beta_0 + \beta_{\text{condition}} x_{\text{condition}} + \beta_{\text{baseline}} x_{\text{baseline}} + \varepsilon \quad (9)$$

In line with modern practice, we assume that this is a single-trial analysis, although the same should hold, albeit less optimally, for aggregated analyses. Including x_{baseline} as a predictor, we use the data to determine the weighting of the baseline correction, with $\beta_{\text{baseline}} = 1$ corresponding to traditional baseline correction and $\beta_{\text{baseline}} = 0$ corresponding to no baseline correction. Now, if the conditions differ in the amount of baseline correction "necessary," we can straightforwardly address this by adding an interaction term to our model:

$$\begin{aligned} y_{\text{window}} = {} & \beta_0 + \beta_{\text{condition}} x_{\text{condition}} + \beta_{\text{baseline}} x_{\text{baseline}} \\ & + \beta_{\text{condition,baseline}} x_{\text{condition}} x_{\text{baseline}} + \varepsilon \end{aligned} \quad (10)$$

This interaction term allows the amount of baseline correction to vary by condition as would be, for example, necessary if changes in the external environment (electrode gel warming up, participant sweating, changes in ambient electrical noise) occur during the course of experiment, especially for block designs. However, even in the case of nonblock designs, this actively accounts for issues resulting from randomization order and can be complemented by added main-effect and interaction terms for the trial sequence (or even smoother terms, cf. H. Baayen, Vasishth, Kliegl, & Bates, 2017; Tremblay & Newman, 2015).

[1]Of course, in a real study, we would probably have multiple predictors, including topographic ones as well as random effects, for example, for by-participant and by-item differences.

As this procedure allows the data to determine how much baseline correction is warranted by condition, it is optimal and not as strongly dependent on the "no systematic differences" assumption. Like GLM-based deconvolution methods, which model mixtures of time-lagged influences on the signal, this technique reduces confounding by explicitly modeling other influences on the signal, instead of mixing them into the response. Moreover, this method includes traditional baseline correction as well as no baseline correction as special cases and thus supersedes those methods. As noted above, this result holds equally well for single-trial time-frequency data under the usual normalization procedure.

The notion of confound is also useful for a more intuitive derivation of the optimality of this approach, where baseline is a covariate. Baseline correction is not there to create a true zero per se but rather as an inferential control (cf. Urbach & Kutas, 2006). As noted previously, good experimental design can and should also function as a way for inferential control, and indeed the usual baseline assumptions correspond exactly to a particular method of experimental control. However, a more general and more powerful technique is to adjust for potential confounds statistically, by including potential confounds as a covariate. Rather than making a priori assumptions about the impact of the confound, this procedure allows for determining its actual influence and allows for a broader class of experimental designs where the confound cannot be controlled via systematic manipulation or experimental procedure (Sassenhagen & Alday, 2016).

This method can also be viewed as a computationally simple special case of regression methods such as rERP (Smith & Kutas, 2014a, 2014b), without lagged predictors and marginalized over distinct time windows. The method presented here has the advantage that it fits much more easily into existing computational and statistical frameworks, trivially works with modern mixed-effects models for simultaneously modeling both participant and item variance (R. H. Baayen, Davidson, & Bates, 2008; Clark, 1973; Judd, Westfall, & Kenny, 2012), and is no more expensive computationally than other contemporary methods (see worked example below). Finally, this method also subsumes and generalizes other baseline-normalization methods such as traditional baseline correction, especially when interactions with the baseline predictor are included.

This method does, however, have a few disadvantages. It functions best with unaggregated (i.e., single-trial) data and explicit regression approaches (i.e., not AN(C)OVA); however, these are considered best practice anyway (for the general statistical preference for explicit estimation, see Cumming, 2014; Kruschke & Liddell, 2017; for insights gleaned from single-trial analyses of ERP data, see, e.g., Frömer, Maier, & Rahman, 2018; Gaspar, Rousselet, & Pernet, 2011; Hauk, Davis, Ford, Pulvermüller, &

Marslen-Wilson, 2006; Pernet, Sajda, & Rousselet, 2011; for the advantages of a multi-level regression approach using mixed-effects models, see R. H. Baayen et al., 2008; Clark, 1973; Judd et al., 2012). Numerically, other issues may arise if there is large signal drift and thus variables on vastly different scales; however, once again best statistical practice, namely, centering and scaling variables, provides a solution to this problem.[2] More challenging is that the additional parameters in these models increase both computational complexity and the amount of data necessary for reliable parameter estimation. This is especially true for models including topographical information (e.g., channel name or position in a multichannel recording). The computational complexity is hard to address, but the requirement for more data is again in line with contemporary best practice to address the chronic lack of power in neuroscience (cf. Button et al., 2013; Szucs & Ioannidis, 2017). Regularization (e.g., ridge regression or LASSO in the frequentist framework, sparsity priors in the Bayesian framework) can also help. This method is also somewhat more difficult to integrate into procedures not based on the GLM, such as independent component analysis (ICA) and source localization, although probably not prohibitively so. For example, this technique would provide an interesting way to improve stationarity and thus potentially enhance IC decompositions of epoched data without depending on the strong filters often used in such contexts.[3] Finally, this method does not completely address issues related to the selection of the baseline interval, which remains an open question, and a researcher degree of freedom, but some general guidelines are suggested in the next section.

5 | RELATIONSHIP TO HIGH-PASS FILTERING AND CHOICE OF BASELINE WINDOW

It is common to refer to baseline correction as an alternative or complementary to (strong) high-pass filtering. However, baseline correction can also be interpreted as a high-pass filter in its own right (albeit an unusual one). In intuitive terms, baseline correction removes the changes in the signal between epochs and can be interpreted as removing slow drifts and thus low-frequency components. Like a filter, baseline correction, both traditional and regression based, also has free parameters that influence its effect on the data.

All things equal, a longer baseline window will tend to be less noisy or variable compared to a shorter one. In statistical terms, a longer baseline window corresponds to a larger sample drawn from a random variable and will thus tend to offer a better estimate of its true mean with less variance (i.e., both more accurate and more precise). However, all things are rarely equal, and longer baseline windows present additional difficulties: they require longer interstimulus intervals (potentially disruptively long ones for many research questions) and/or potentially include parts of the evoked response from the previous stimulus, thus changing the meaning of reference point for later evoked potentials. This suggests that a baseline window on the order of a few hundred milliseconds may be the sweet spot for many experimental designs under typical laboratory conditions without large high-frequency artifacts (see empirical example below for a brief comparison of different baseline windows).

Beyond the length of the baseline window, the relative position of the baseline window to the time-locking events and critical events is also important. Because the position of the baseline window within an epoch is absolute and not relative compared to a given sample (as in a typical filter), baseline correction will generally not remove slow drifts within an epoch. In traditional baseline correction, the entire epoch is shifted by a constant offset, and thus the overall slope is not affected: translations are shape-preserving transforms. In the regression-based correction proposed here, the drift away from the calibration given by baseline will eventually lead to the baseline weight shrinking to zero. This is unsurprising in the sense that a distant baseline window is a poor baseline window (e.g., the first 2 s of EEG recording are not used as the baseline window for all trials in that recording). As such, baseline correction is not a substitute for but rather a complement to traditional high-pass filtering.

The choice of baseline window should also be shaped by the research question. The logic of baseline correction, as highlighted by Urbach and Kutas (2006), is not to establish a true zero (which may or may not be meaningful for a measure such as voltage that is inherently a difference) but rather a meaningful reference point or control with which to compare successive changes and thereby infer causality. For a classical prestimulus baseline, the ERP thus shows the change in the electric field following the stimulus (or, more generally, event of interest): the state after the stimulus relative to the (average) state before. For a baseline consisting of the entire epoch, the ERPs show the change in the electric field relative to its average of a time interval that includes the event of interest. This does not show as directly that the state afterward is different than the state before and instead only shows the difference to the average state. When the difference to the average state is larger after an event of interest than the difference to the

[2]This is sometimes addressed as part of the signal processing, via a special case of baseline correction, namely, subtracting the mean of the whole trial from each trial. However, as the activity between conditions is assumed to differ between trials, this again violates the assumptions of baseline correction and can introduce effects into other time windows. This is especially problematic for large-amplitude and/or prolonged effects. While not problematic for statistical analyses carefully focused on a single time window of interest, this is still less optimal than simply scaling the variables in the regression model.

[3]We are indebted to a helpful reviewer for suggesting this approach.

average state before the event of interest, then this can also be taken as indirect evidence of event-related change; however, this second stage of "difference of differences" is implicitly an additional baseline correction to the prestimulus interval. The no-baseline-correction case corresponds to assuming that the reference point aligns with true zero, which may be a reasonable assumption, for example, for studies with stronger high-pass filtering and comparable stimulation before the critical event. The advantage of using regression-weighted baseline correction is that the data determine the evidence that the chosen reference point (baseline window) differs from true zero and how to weight its contribution because the reference point itself is a noisy measurement. In other words, using a deterministic baseline is ignoring the error bars on the control given by the baseline window.

This is crucial when interpreting topographies. Traditional baseline correction necessarily projects the inverse scalp topography into the epoch, but the weighting in the regression-based approach properly controls for scalp topography instead of forcibly shifting it. This is achieved in two ways. First, the weighting of the baseline window can differ by electrodes. Second, the weighting of the baseline window can differ by condition. In either case, this can be achieved by performing the baseline correction on each electrode or condition separately (as in traditional baseline correction) or by including topographical position or condition as interaction effects in the regression model (for a pooled estimate). By applying such proper statistical control, we can avoid many of the biases that lie at the heart of Urbach and Kutas's (2006) arguments. The empirical example in the next section shows how traditional baseline correction can be misleading in such cases, but the regression-based approach properly controls for topographical differences in the baseline conditions.

In brief, baseline correction serves a similar role to high-pass filtering and suffers many of the same potential pitfalls in terms of artifacts, both causal and acausal. Moreover, each has a number of similar tradeoffs: longer baseline intervals and stronger high-pass filters better correct for some types of noise in the signal but at increased risk of additional artifacts. However, one does not completely replace the other, and the combined choice of baseline window and high-pass filter should reflect the tradeoffs necessary for a particular experimental design. Regression-based baseline correction supersedes traditional baseline correction but does not eliminate the need for appropriate high-pass filtering.

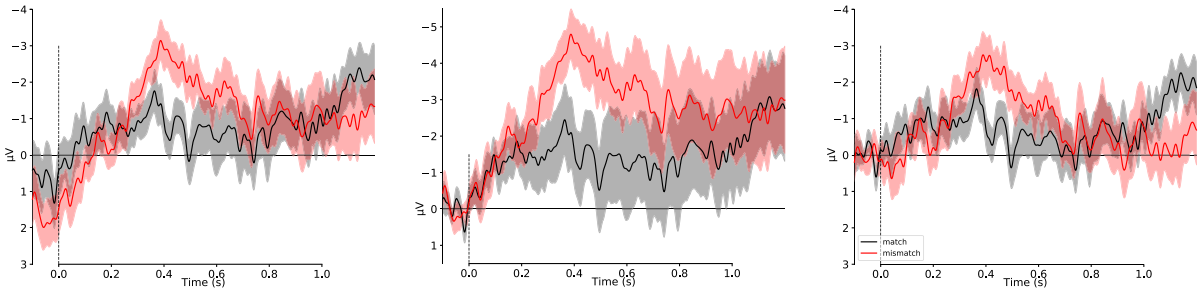
6 | EMPIRICAL EXAMPLE: N400 PARADIGM WITH ENVIRONMENTAL NOISE

In the following, we aim to demonstrate the claims above via an empirical example. We reanalyze data from Tromp, Peeters, Meyer, and Hagoort (2017), a classical semantic mismatch N400 paradigm, but conducted in virtual reality with a cross-modal mismatch. The virtual reality setting presents a particular challenge because of the potential for environmental noise and movement artifacts. Such noise and artifacts could potentially cause signal changes despite no violation of the "no systematic differences in electrophysiological activity" assumption and thus necessitate a correction for signal drift.

Using MNE-Python v0.17.1 (Gramfort et al., 2013) and in line with the original analysis, (continuous, nonepoched) data were rereferenced to the linked mastoids and band-pass filtered from 0.1 to 40 Hz (pass-band edge; zero-phase FIR filter with a Hamming window and `fir_design = firwin`, all other parameters left as `auto`). These filter settings should eliminate line noise and very slow drifts without inducing problematic artifacts, but traditional wisdom suggests that they do not eliminate the need for baseline correction (cf. Tanner et al., 2016). As in the original analysis, the baseline interval consisted of the 100 ms immediately before (auditory) stimulus onset. Analyses conducted with other high-pass filter edges (0.1, 0.3, 0.5, 1.0 Hz) are presented below in summary form for comparison but are not discussed at depth nor further analyzed. Trials with instantaneous amplitude exceeding ±75 μV were excluded from further analysis. Although Tromp and colleagues (2017) analyzed both the N400 time window and an earlier time window, we restrict ourselves to their N400 window (350–600 ms).

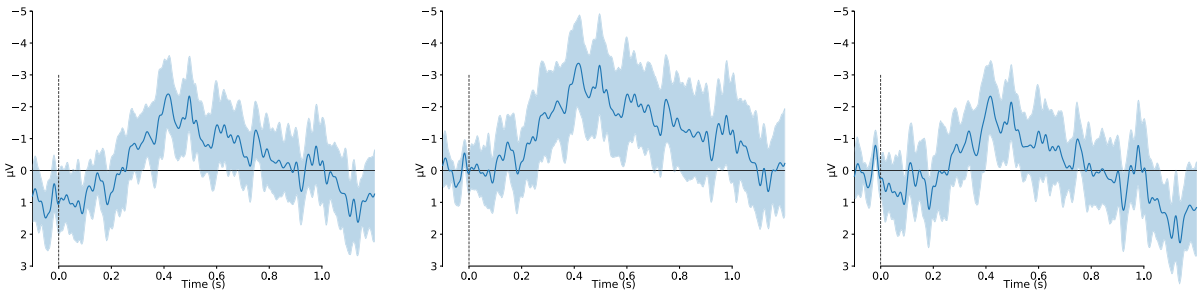
It is important to note that the original data were recorded at 500 Hz and filtered online with a low-pass filter at 200 Hz and a high-pass filter at 0.016 Hz. The file metadata show that the high-pass filter was applied both in hardware and in software, while the low-pass filter was applied only in software. Although Tromp and colleagues (2017) originally reported online high-pass filtering at 0.01 Hz, Brain Products amplifiers specify their cutoff in time (here: 10 s), and the corresponding frequency cutoff is calculated as $1/2\pi t$ following analog filter convention and not as $1/t$ as is common in other areas. As such, the raw data already reflect two forward passes of a weak high-pass filter. This will, of course, greatly attenuate the sorts of drift that baseline correction serves to correct but is not an unusual recording setup and as such demonstrates that the role baseline correction plays under actual laboratory conditions.

All analysis source code as well as the preprocessed single-trial data are available on OpenScience Framework (https://osf.io/pnaku/). There are data for each of the above filter settings as well as for several different baseline windows (500 ms prestimulus, 200 ms prestimulus, 100 ms prestimulus, 200 ms poststimulus, average across entire epoch). It is beyond the scope of this manuscript to discuss all possible combinations of baseline interval and filter settings in depth, but we do briefly examine the impact of the baseline interval for the primary high-pass filter setting (0.1 Hz) below.



(a) No baseline correction **(b)** Traditional baseline **(c)** Regression-based baseline

FIGURE 1 Comparison of baseline-correction strategies for waveforms at the apex electrode (Cz)



(a) No baseline correction **(b)** Traditional baseline **(c)** Regression-based baseline

FIGURE 2 Comparison of baseline-correction strategies for difference waves at the apex electrode (Cz)

6.1 | Differences in the baseline are illusory

The grand-averaged waveforms from the apex electrode Cz are presented in Figure 1 with the corresponding difference waves in Figure 2. Figure 1a presents the waveform without any baseline correction. Although the grand averages themselves look distinct in the baseline window, we see that the confidence intervals overlap, and correspondingly the confidence interval for the difference wave crosses zero (Figure 2a). We are thus unable to reject the null hypothesis that the observed difference between conditions in the baseline interval occurred by chance alone. Less rigorously, the waveforms are statistically indistinguishable in the baseline window, and the apparent differences in the baseline window are not distinguishable from noise. Figure 1b presents the waveform with traditional baseline correction. We note how the confidence intervals become broader; moreover, there is an apparent, yet misleading, prolonged separation of the waveforms well beyond the N400 time window, which is also apparent in the difference wave in Figure 2b. Again, the overlap in the confidence intervals suggests that this difference is not distinguishable from noise; however, this distinction would be lost in typical ERP plots without confidence intervals. Finally, Figure 1c presents the regression-based baseline strategy applied to each time point. We see that the confidence intervals are much narrower than in the traditional baseline correction. Moreover, the overall time course of the N400 effect is much more apparent and much more temporally focal in the difference plot (Figure 2c).

Figure 3 displays the topography of the grand-averaged difference waves. Note that the overall topography does not change greatly between baseline-correction methods for this experiment because stimulation before onset of the critical item was comparable. Taken together, Figures 2 and 3 suggest that regression-based baseline correction reduces the size of the N400 effect. This is not quite accurate; instead, traditional baseline correction leads to a slight overestimation of the size of the N400 effect. This is discussed in more depth below.

Most interestingly, the later N400 effect around 600–800 ms reported by Tromp and colleagues (2017) with traditional baseline correction has a different topography than the early one around 300–500 ms.[4] While they reported no overall interaction between condition and topography within each time window, they did not compare topographies between time windows. Meanwhile, both no baseline correction and regression-based baseline correction suggest an extremely weak effect near zero across the entire scalp (Figure 3a,c). Examining the topography in the baseline window given in the no baseline–correction plot

[4]We are indebted to a helpful reviewer for pointing out this shift in topography and positing that it may be a baseline artifact.

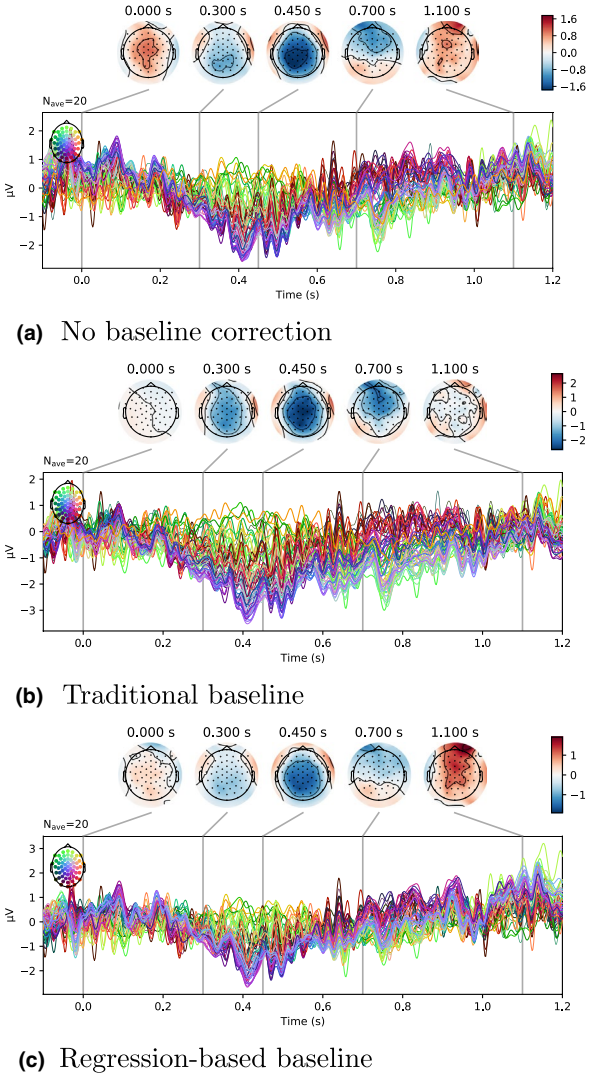


FIGURE 3 Comparison of baseline-correction strategies for the topography of difference waves

(Figure 3a), we see that traditional baseline correction projects this small, albeit nonsignificant, difference in means forward in time, where it combines with the minimal effect present in that time window to generate the larger observed effect with its distorted topography and duration. This graphical impression is supported by analysis with mixed-effects models: the inclusion of the baseline in the model improves fit and removes the effect of condition (Figure 4). In contrast, the primary N400 effect in the 350–600 ms time window retains its topography across correction methods.

For all of these plots, we note that the bootstrap confidence intervals computed samplewise per electrode on single-subject averages do not correspond directly to the statistics used in the analysis below. In particular, the analyses below include subject and item variance simultaneously and are computed on trialwise window and region of interest (ROI) averages. The window and ROI averaging will generally increase the signal-to-noise ratio, and as has been noted many times (e.g., R. H. Baayen et al., 2008; Clark, 1973; Judd et al., 2012), item variance cannot be ignored, especially in language studies. This is apparent (see Table 1) where the between-item variance is larger than the between-subject variance.

6.2 | Prestimulus baseline influence on later components is not what you think

The misleading duration and amplitude of the N400 effect in the plot with traditional baseline correction is partly the result of traditional baseline correction's ability to bias later components in the wrong direction. Figure 5 shows the correlation of the ERP for each condition with the baseline

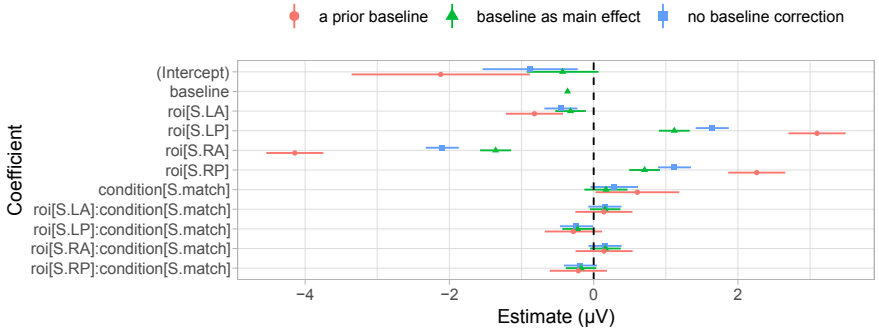


FIGURE 4 Coefficient plot comparing estimates from different baseline correction strategies in the later time window (600–800 ms). Intervals are 95% profile confidence intervals. Note the extremely small, yet extremely precise, estimate for the (effect of the) baseline window. Model selection preferred the regression model including baseline as a main effect but not further interacting with ROI or condition; see below for a more extensive example using the primary window of interest (350–600 ms). The impact of the small bias introduced by traditional baseline correction is apparent in the confidence interval for the effect of condition—even a small change downward would have led to a rejection of the null hypothesis

TABLE 1 Summary of full model for the primary window of interest (350-600 ms) with pairwise interactions between topography, manipulation, and the baseline

Linear mixed model fit by maximum likelihood				
AIC	BIC	logLik	Deviance	df.resid
40623	40782	−20289	40577	7187
Min	1Q	Median	3Q	Max
−5.44	0.64	0.01	0.64	4.42
Random effects				
Groups	**Term**	***SD***	**Corr**	
Item	(Intercept)	1.02305		
	condition[S.match]	0.68514	−0.366	
Subject	(Intercept)	0.48629		
	condition[S.match]	0.67955	0.228	
Residual		3.94258		
Number of observations: 7210, Groups: item, 80; subject, 20.				
Fixed effects				
	Estimate	***SE***	***t* value**	
(Intercept)	−0.92	0.17	−5.6	
baseline	−0.2	0.0088	−23	
roi[S.LA]	0.24	0.096	2.5	
roi[S.LP]	0.19	0.094	2	
roi[S.RA]	−0.23	0.1	−2.2	
roi[S.RP]	−0.11	0.094	−1.2	
condition[S.match]	0.47	0.18	2.7	
baseline:roi[S.LA]	0.011	0.016	0.7	
baseline:roi[S.LP]	−0.014	0.018	−0.77	
baseline:roi[S.RA]	0.0039	0.015	0.25	
baseline:roi[S.RP]	−0.01	0.018	−0.57	
baseline:condition[S.match]	−0.033	0.0087	−3.8	
roi[S.LA]:condition[S.match]	−0.11	0.093	−1.2	
roi[S.LP]:condition[S.match]	0.11	0.094	1.2	
roi[S.RA]:condition[S.match]	−0.1	0.095	−1.1	
roi[S.RP]:condition[S.match]	0.067	0.094	0.71	

Note: All categorical contrasts are sum coded. ROIs are named by laterality (L vs. R) and sagittality (A vs. P) or the midline (M). Model fitted with *lme4* version 1.1.20 (Bates, Maechler, Bolker, & Walker, 2015).

interval over time. Unsurprisingly, the correlation with the mean of the baseline interval is quite high within the baseline interval; however, the correlation drops off rapidly, reaching zero less than 200 ms after stimulus onset. Somewhat disconcertingly, the correlation in typical N400 and P600 time windows is nonzero and negative. This suggests that traditional baseline correction is shifting the waveform in the wrong direction. As voltage is inherently a relative measure, this bias, shared among all conditions, is not particularly problematic per se. Nonetheless, the low magnitude of the correlation at larger latencies indicates that there is little shared covariance between the baseline window and the target window. In other words, applying traditional a priori baseline correction fails to correct bias introduced by the baseline. Moreover, traditional baseline correction may introduce additional bias and will necessarily introduce the additional variance from the baseline interval. This suggests that the baseline interval is most relevant for the early exogenous, perceptual components. Again, including the baseline interval as a predictor in the statistical model applies the correct amount of correction as determined by the data—and that level of correction is expected to differ between data sets. For example, DC recordings without online or offline high-pass filtering will necessarily require more correction than those such as here with both online and offline high-pass filters.

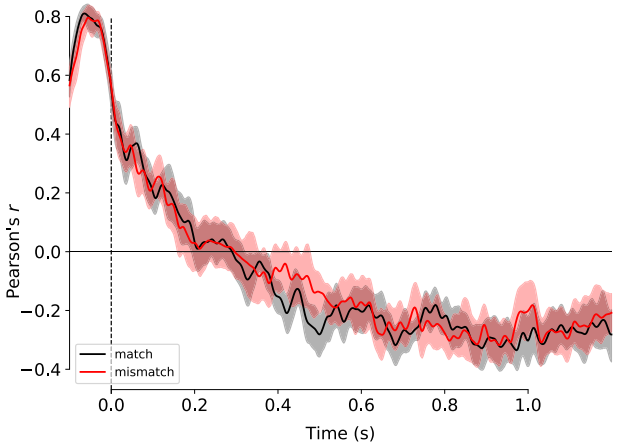


FIGURE 5 Correlation of electrophysiological signal with baseline interval at the apex electrode (Cz)

We see this in the mixed-effects model for the N400 time window, presented in Table 1 and Figure 6. Although the main effect for the baseline window has a large *t* value, the actual size of the effect is quite small and in the wrong direction. (Recall from above that traditional baseline correction corresponds to a regression weight of +1). We also note that there is an interaction of baseline with condition, which traditional baseline correction could not have accommodated.

6.3 | Model complexity and fit and their impact on statistical power

While the model presented in Table 1 may seem much more complex to fit and interpret than a model without the baseline predictors, this is not the case. As elsewhere in statistics, we can include additional covariates as controls without further interpreting those covariates. In other words, we can safely ignore the terms related to baseline correction, but we cannot omit them from the model. As reflected in the shifted vertical midpoint in Figure 1, the baseline term will have an impact if we compute, for example, marginal means, but that does not preclude us from interpreting the effects attributable purely to our experimental manipulation. Moreover, if the interpretation of the interaction between the baseline and the experimental manipulation is of interest, then it is no different than the interpretation of the interaction between topographical predictors and the experimental manipulation.

Following the ongoing debate about the tradeoffs in Type I error, power, and model complexity (e.g., Barr, Levy, Scheepers, & Tily, 2013; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017), we can consider the impact of additional predictors on model fit and statistical power. Figure 7 shows that the improved fit resulting from including the baseline window as a predictor more than compensates for the potential loss in power from the additional predictors (power estimated using the simr package, Green & MacLeod 2016). Moreover, the reduced variance in the dependent variable results in faster convergence of the numerical optimization procedure, and thus computation time is also not worsened by the additional model complexity. For this particular data set, the models with additional terms for the interaction of the baseline with condition and topography do show an improved fit (as measured by log likelihood), but the accompanying increase in model complexity exceeds the corresponding improvement

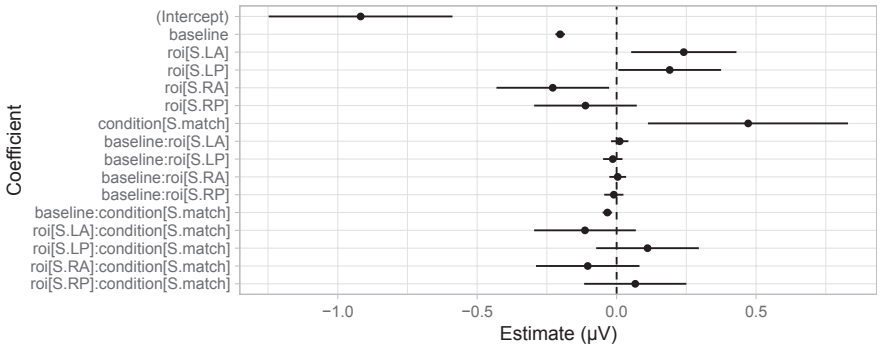


FIGURE 6 Coefficient plot for the model presented in Table 1. Intervals are 95% profile confidence intervals. Note the extremely small, yet extremely precise, estimate for the (effect of the) baseline window

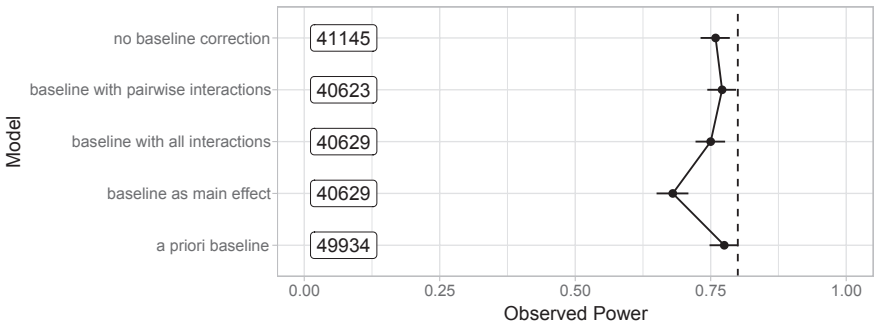


FIGURE 7 Statistical power and model fit for different types of baseline correction

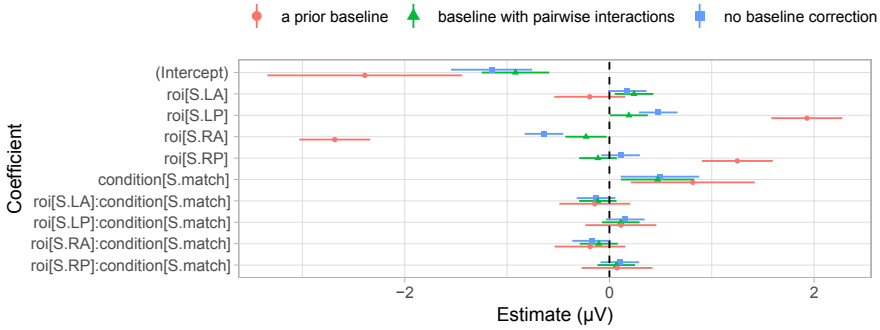


FIGURE 8 Coefficient plot comparing the estimates from the models corresponding to different baseline strategies. Intervals are 95% profile confidence intervals. Note the much larger confidence intervals for the a priori baseline but otherwise overall similar pattern of effects for the experimental manipulation and its topography. The differences in main effects in topography are an example of the topographical biases inherent in traditional baseline correction and reflect the combined topography of the baseline interval and average topography across both conditions, while the interaction model separates these effects

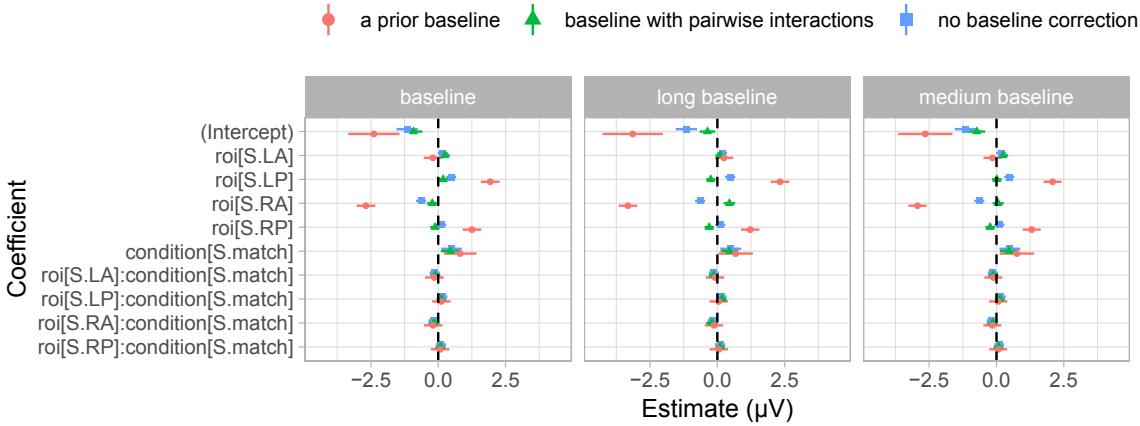


FIGURE 9 Coefficient plot comparing the estimates from the models corresponding to different baseline strategies with different baseline windows. Intervals are 95% profile confidence intervals. The long baseline corresponds to 500 ms prestimulus, the medium to 200 ms prestimulus, and the "default" baseline to 100 ms prestimulus. The different baseline strategies and windows were estimated with separate models

in model fit when comparing the pairwise interaction model to the full interaction model ($\Delta AIC = 6$, and corresponding likelihood-ratio test $\chi^2(4) = 2.8$, $p = .6$). We therefore prefer the more parsimonious pairwise interaction model over the full interaction model. Crucially, the model with traditional, a priori baseline correction performs the worst in terms of model fit. The minimal apparent increase in power is thus irrelevant because a poorly fitting model calls the overall validity of inference into question. We see here empirically what we demonstrated mathematically above: traditional baseline correction reduces power and biases our inferences.

For comparison, the estimates from a priori baseline, no baseline, and the pairwise model are plotted in Figure 8 (see also Figure 4 for a similar comparison in the late N400 time window examined in a post hoc analysis by Tromp and colleagues). Overall, the pattern of effects is similar across models, except that the model with the a priori baseline has much larger estimates and larger confidence intervals. For the main effects of topography, this reflects the topographical biases inherent in traditional baseline correction and reflects the combined topography of the baseline interval and average topography across both conditions, while the interaction model separates these effects.

The larger estimate for the experimental manipulation also leads to its high power estimate (cf. Figure 7). Although its confidence interval is much broader than the other models, the mean value is higher and so the lower edge of the confidence is further away from zero. This in turns leads to higher observed power, which is known to be biased in this way (cf. Gelman & Carlin, 2014; Hoenig & Heisey, 2001).

6.4 | Choice of baseline window and high-pass filter

The baseline window and high-pass filter used in the analysis thus far were chosen to match the original analysis by Tromp and colleagues (2017). Given the overall experimental design and considerations on the impact of baseline window discussed above (Section 5), we do not expect a large difference for longer prestimulus windows. We tested

this empirically by computing the same pairwise interaction model for the same baseline (100 ms prestimulus), a long baseline (500 ms prestimulus), and a medium baseline (200 ms prestimulus). As Figure 9 shows, the overall pattern of effects, both between and within models, did not change between conditions, although the absolute magnitude of the intercept term (reflecting the average voltage across all conditions and ROIs) did change. Similarly, the weight awarded to the baseline window changed (see Figure 10), but its interactions with ROIs and condition did not (reflecting an overall matching of the baseline prestimulus interval across conditions). The change in both the intercept and weight of the baseline term reflects a change in the absolute voltage measured in the N400 window, but the absolute change on a relative scale is less interesting than the impact it has on the estimate of the effect of interest, which was minimal: the estimates for condition and its topographical interactions did not differ much between baseline windows. Note that different experimental designs with different stimulation and noise constraints can lead to a longer or shorter baseline being

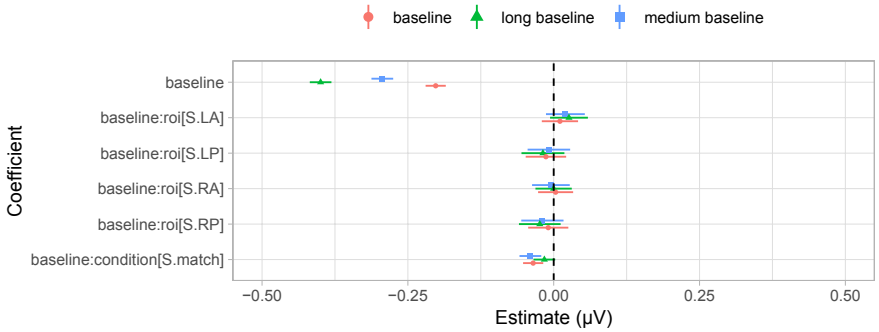


FIGURE 10 Coefficient plot comparing the estimated weights awarded to different baseline intervals. Intervals are 95% profile confidence intervals. The long baseline corresponds to 500 ms prestimulus, the medium to 200 ms prestimulus, and the "default" baseline to 100 ms prestimulus. The different baseline windows were estimated with separate models, all including pairwise interactions of the baseline interval and other predictors

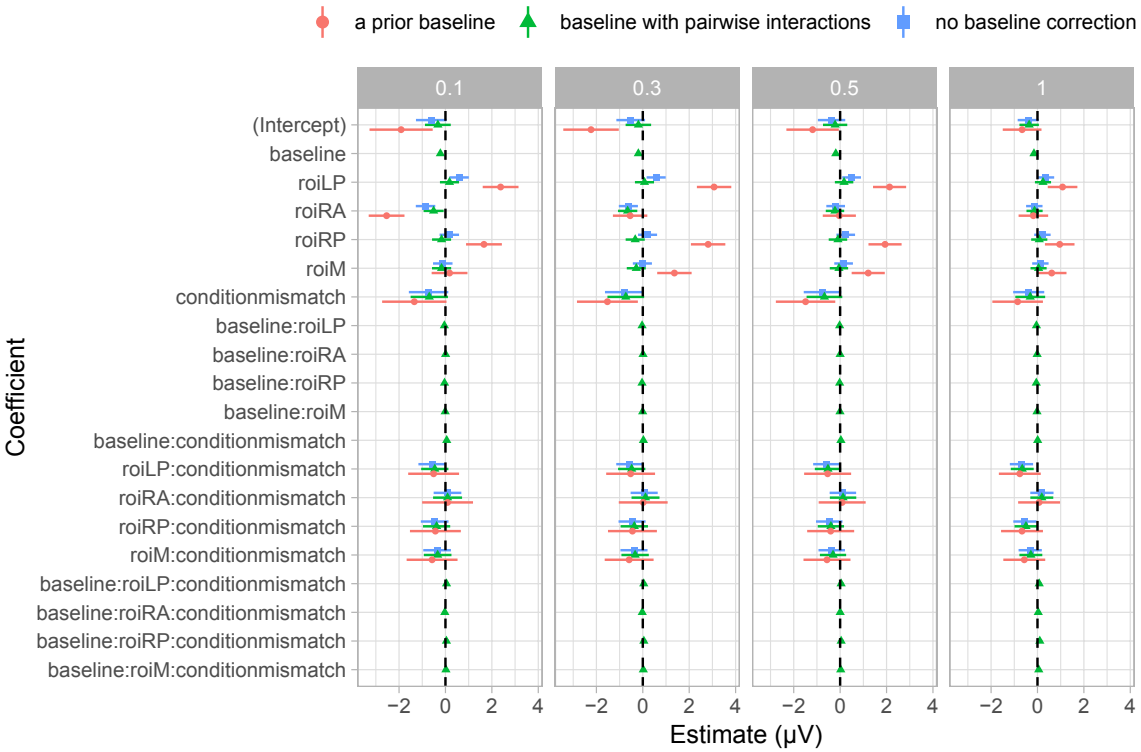


FIGURE 11 Coefficient plot comparing the estimates from the models corresponding to different baseline strategies and different high-pass filter settings. Intervals are 95% profile confidence intervals. All filters are band-pass zero-phase FIR filters with an upper pass-band edge of 40 Hz and a lower pass-band edge corresponding to the value in the plot. The different baseline strategies and filter settings were estimated with separate models. Note that stronger filtering shrinks all effects to zero but those attributable to drift (intercept, baseline, topographical main effects) and not the targeted experimental manipulation (condition and its interactions) more strongly

preferable, but for this experiment, the choice of prestimulus baseline window did not have a huge impact.

Similarly, the choice of high-pass filter did not greatly impact the effect of interest here, as seen in Figure 11. All filter settings with the exception of the relevant pass-band edge were the same as above (zero-phase FIR), and for simplicity these models were computed using only the 100-ms prestimulus window as the baseline correction. Stronger filtering shrinks all effects toward zero but those attributable to drift (intercept, baseline, topographical main effects) more strongly than the fast changes due to targeted experimental manipulation (condition and its interactions). In addition to the potential to shrink events of interest to zero with strong enough filtering, filters can also introduce other artifacts not obvious in the statistical models, as discussed at length in the Tanner-Maess debate. Moreover, pass-band edge is not the only relevant filter setting—the choice of causal versus acausal, zero-phase or not, filter-length and IIR versus FIR—all involve a number of tradeoffs whose scope exceeds the present manuscript.

6.5 | Bayesian analysis

Despite the theoretical and empirical evidence presented above, some researchers may still have a strong a priori belief in the necessity of the traditional baseline procedure. To that end, we again note that the data-driven, model-based approach presented here will yield traditional baseline correction, when the data support it. Moreover, we can accommodate our a priori beliefs as part of the statistical model. Using the R package *brms* (Bürkner, 2017) to interface with the probabilistic programming language Stan (Carpenter et al., 2017, RStan version 2.18.2), we also ran a Bayesian analysis with a main effect of baseline interval and main effects of and interactions between experimental condition and scalp topography. For the baseline interval, we used a Student's t prior with three degrees of freedom, centered at +1 and variance equal to 0.001. This leads to a very sharp spike centered at 1 with heavy tails—in more casual terms, this is a very strong belief in traditional baseline correction with nonetheless a willingness to change given enough evidence. For the condition and topographical factors, we used normal priors centered at 0 and with standard deviation equal to 2. This is equivalent to the assumption that 60% of effects are smaller than ±2 μV and 95% effects are smaller than ±4 μV, which is a reasonable "no outrageous" effects assumption for language-related ERPs.

Figure 12 presents the resultant change in beliefs about the correct weighting for the baseline interval. Even starting from such a strong assumption, the posterior distribution still clearly places the most credibility on a small, yet nonzero weighting for the baseline interval in the direction opposite the traditional direction.

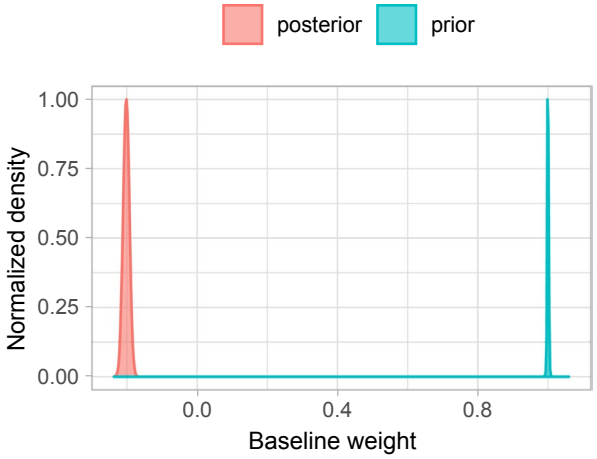


FIGURE 12 Bayesian analysis of the appropriate weighting of the baseline interval

7 | CONCLUSION

Baseline correction is in many ways the twin of filtering in EEG preprocessing, serving both to replace stronger filtering and ultimately functioning as a filter itself (see above discussion in *Psychophysiology* and *Journal of Neuroscience Methods*). However, traditional baseline correction is self-defeating, increasing noise, and not affecting signal in exactly those situations fulfilling its assumptions. Here, we have presented a straightforward extension of the modern statistical analysis that supercedes the traditional baseline correction, allowing the data to dynamically determine the strength of the correction, while including both traditional baseline correction and no baseline correction as limiting cases. Extending Tanner and colleagues' (2016) comments a bit, we can find out whether and how much baseline correction is a good idea.

ACKNOWLEDGMENTS

Johanna Tromp and David Peeters kindly shared their data with me for the practical example. I am grateful to several colleagues who have commented on various drafts of this proposal in both casual conversation and written form, including but not limited to Ingmar Brilmayer, Franziska Kretzschmar, Benedikt Ehinger, Florian Hintz, Greta Kaufeld, Suzanne Jongman, Darren Tanner, and Andreas Widmann. In particular, I am thankful to Andrea Martin for suggesting the current (sub)title. Two anonymous reviewers provided invaluable feedback on theoretical and practical aspects. All remaining mistakes are my own.

ORCID

Phillip M. Alday https://orcid.org/0000-0002-9984-5745

REFERENCES

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234. https://doi.org/10.1016/j.jml.2016.11.006

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2015). Fitting linear mixed-effects models using Lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, S. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *20*(10), 1–23. https://doi.org/10.1177/0956797613504966

Frömer, R., Maier, M., & Rahman, R. A. (2018). Group-level EEG-processing pipeline for flexible single trial-based analyses including linear mixed models. *Frontiers in Neuroscience*, *12*, 48. https://doi.org/10.3389/fnins.2018.00048

Gaspar, C. M., Rousselet, G. A., & Pernet, C. R. (2011). Reliability of ERP and single-trial analyses. *NeuroImage*, *58*(2), 620–629. https://doi.org/10.1016/j.neuroimage.2011.06.052

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., … Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*, 267. https://doi.org/10.3389/fnins.2013.00267

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *NeuroImage*, *30*(4), 1383–1400. https://doi.org/10.1016/j.neuroimage.2005.11.048

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician*, *55*(1), 19–24.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *1–29*, https://doi.org/10.3758/s13423-016-1221-4

Lau, E., Stroud, C., Plesch, S., & Phillips, C. (2006). The role of structural prediction in rapid syntactic analysis. *Brain Language*, *98*(1), 74–88. https://doi.org/10.1016/j.bandl.2006.02.003

Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.

Maess, B., Schröger, E., & Widmann, A. (2016a). High-pass filters and baseline correction in M/EEG analysis—Continued discussion. *Journal of Neuroscience Methods*, *266*, 171–172. https://doi.org/10.1016/j.jneumeth.2016.01.016

Maess, B., Schröger, E., & Widmann, A. (2016b). High-pass filters and baseline correction in M/EEG analysis. Commentary on: How inappropriate high-pass filters can produce artefacts and incorrect conclusions in ERP studies of language and cognition. *Journal of Neuroscience Methods*, *266*, 164–165. https://doi.org/10.1016/j.jneumeth.2015.12.003

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Pernet, C. R., Sajda, P., & Rousselet, G. A. (2011). Single-trial analyses: Why bother? *Frontiers in Psychology*, *2*, 322. https://doi.org/10.3389/fpsyg.2011.00322

Sassenhagen, J., & Alday, P. M. (2016). A common misapplication of statistical inference: Nuisance control with null-hypothesis significance tests. *Brain and Language*, *162*, 42–45. https://doi.org/10.1016/j.bandl.2016.08.001

Smith, N. J., & Kutas, M. (2014a). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*, *52*, 157–168. https://doi.org/10.1111/psyp.12317

Smith, N. J., & Kutas, M. (2014b). Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. *Psychophysiology*, *52*, 169–181. https://doi.org/10.1111/psyp.12320

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*, 3. https://doi.org/10.1371/journal.pbio.2000797

Tanner, D., Morgan-Short, K., & Luck, S. J. (2015). How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology*, *52*(8), 997–1009. https://doi.org/10.1111/psyp.12437

Tanner, D., Norton, J. J. S., Morgan-Short, K., & Luck, S. J. (2016). On high-pass filter artifacts (they're real) and baseline correction (it's a good idea) in ERP/ERMF Analysis. *Journal of Neuroscience Methods*, *266*, 166–170. https://doi.org/10.1016/j.jneumeth.2016.01.002

Tremblay, A., & Newman, A. J. (2015). Modeling nonlinear relationships in ERP data using mixed-effects regression with R examples. *Psychophysiology*, *52*, 124–139. https://doi.org/10.1111/psyp.12299

Tromp, J., Peeters, D., Meyer, A. S., & Hagoort, P. (2017). The combined use of virtual reality and EEG to study language processing in naturalistic environments. *Behavior Research Methods*, *50*(2), 862–869. https://doi.org/10.3758/s13428-017-0911-9

Urbach, T. P., & Kutas, M. (2006). Interpreting event-related brain potential (ERP) distributions: Implications of baseline potentials and variability with application to amplitude normalization by vector scaling. *Biological Psychology*, *72*(3), 333–343. https://doi.org/10.1016/j.biopsycho.2005.11.012

Widmann, A., Schröger, E., & Maess, B. (2015). Digital filter design for electrophysiological data—A practical approach. *Journal of Neuroscience Methods*, *250*, 34–46. https://doi.org/10.1016/j.jneumeth.2014.08.002

How to cite this article: Alday PM. How much baseline correction do we need in ERP research? Extended GLM model can replace baseline correction while lifting its limits. *Psychophysiology*. 2019;56:e13451. https://doi.org/10.1111/psyp.13451