Away from arbitrary thresholds: using robust statistics to improve artifact rejection in ERP

Phillip M. Alday

Jeroen van Paridon

October 2020

Abstract

Traditionally, artifacts are handled one of two ways in ERP studies: (1) rejection of affected segments and (2) correction via e.g. ICA. Threshold-based rejection is problematic because of the arbitrariness of the chosen limits and particular threshold criterion (e.g. peak-to-peak, absolute, slope, etc.), resulting in large researcher degrees of freedom. Manual rejection may suffer from low inter-rater reliability and is often done without appropriate blinding. Additionally, rejections are typically done for an entire trial, even if the ERP measure of interest isn't impacted by the artifact in question (e.g. motion artifact at the end of the trial). Additionally, fixed thresholds cannot distinguish between non-artifactual extreme values (i.e. those arising from brain activity and which have some 'signal' and some 'noise') and truly artifactual values (e.g. those arising from muscle activity or the electrical environment and which are essentially pure 'noise'). These aspects all become particularly problematic when analyzing EEG recorded under more naturalistic conditions, such as free dialogue in hyperscanning or virtual reality. By using modern, robust statistical methods, we can avoid setting arbitrary thresholds and allow the statistical model to extract the signal from the noise. To demonstrate this, we re-analyzed data from a multimodal virtual-reality N400 paradigm. We created two versions of the dataset, one using traditional threshold-based peak-to-peak artifact rejection (150µV), and one without artifact rejection, and examined the mean voltage at 250-350ms after stimulus onset. We then analyzed the data with both robust and traditional techniques from both a frequentist and Bayesian perspective. The non-robust models yielded different effect estimates when fit to dirty data than when fit to cleaned data, as well as different estimates of the residual variation. The robust models meanwhile estimated similar effect sizes for the dirty and cleaned data, with slightly different estimates of the residual variation. In other words, the robust model worked equally well with or without artifact rejection and did not require setting any arbitrary thresholds. Conversely, the standard, non-robust model was sensitive to the degree of data cleaning. This suggests that robust methods should become the standard in ERP analysis, regardless of data cleaning procedure.

Introduction

The detection and removal of artifacts (either through trial/segment rejection or via correction) plays a large role in the analysis of event-related potentials (ERPs). Introductory texts devote entire chapters to an overview of the sources of common artifacts and standard semi-automatic *rejection* techniques (cf. Luck 2005) and still require online supplementary materials to adequately address modern artifact *correction* techniques such as independent-component analysis (ICA; Jung et al. 1997). Despite all the attention that has been devoted to the topic of artifacts, there seems to be little consensus as to which rejection or correction techniques (and corresponding thresholds or algorithmic parameters) are optimal, with manual rejection still often seen as the gold standard (cf. Luck 2005), but ICA gaining broader acceptance in the coregistration (between EEG and eye-tracking) community (cf. Dimigen 2020). At present, artifact rejection and correction remain a source of researcher degrees of freedom (e.g. in manual artifact rejection) and a reason that many ERP studies are not easily reproducible (Simmons, Nelson, and Simonsohn 2011; Gelman and Loken 2013).

The traditional approach to artifact correction or rejection focuses on the taxonomy of their origin (e.g. drift, eye movements, muscle activity) in order to develop methods to detect artifacts, but ultimately this is a distraction; what matters is the impact artifacts have, collectively, on the statistical and inferential procedure.

From a statistical perspective, artifacts are simply a source of *noise*, and it is the job of statistical models to separate the signal from the noise. However, classical statistical approaches are sensitive to distributional assumptions about noise, with the relative proportion of outliers to inliers being of particular concern. In technical terms, classical approaches often have a low *breakdown point*, i.e. the smallest proportion of outliers in the observations that can result in the estimate being arbitrarily large or small (Wilcox 2010). For example, the mean has a finite-sample breakdown point of $1/n$ because a single (sufficiently extreme) outlier can result in an estimate being arbitrarily far from the true value, while the median has a finite-sample breakdown point of approximately $1/2$ because roughly one half of all sampled values must be outliers for the median to be arbitrarily far from the true value. In other words, individual data points can have a large influence on the estimate of the mean, but not on the estimate of the median. This fragility of mean estimates is particularly problematic because the mean is the foundation of most classical statistical techniques (including t-tests, ANOVA, standard linear regression and mixed-effects regression). When using classical statistical methods, the traditional approach of identifying and removing artifacts is therefore prudent: Outliers and their problematic influence on effect estimates are minimized by removing known sources of noise, thereby increasing the signal-to-noise ratio (SNR) and the ability of the statistics to extract the signal from the remaining noise. However, a traditional approach to artifact identification and rejection based on arbitrary thresholds will invariably suffer from one of two problems: Either the researcher is overzealous in rejecting artifacts, at which point the cleaned data will contain only a portion of the actually informative observations in the dataset, leading to worse inference both in terms of the estimated effect size and the uncertainty in the effect size estimate, or the researcher is too lenient in rejecting artifacts, leaving some portion of outlier observations in the dataset, which when using classical, mean-based statistical models results in biased effect size estimates.

If we want to avoid the problems inherent in traditional artifact rejection and classical statistics, we need to use statistical models that have a high breakdown point (like the median) instead of a low breakdown point (like the mean). These statistical techniques are less fragile than classical techniques and therefore they are collectively known as *robust statistics* (cf. Wilcox 2010, 2012). As the name indicates, robust statistics are robust to violations of their assumptions, most importantly assumptions regarding the distribution of outliers. Modern robust approaches can substantially improve statistical power and accuracy when the assumptions of classical approaches are not met, while performing nearly as well as classical approaches when their (i.e. classical approaches') assumptions *are* met. In other words, robust statistics provide a principled statistical approach to dealing with outliers and extreme values, including EEG artifacts. As such, robust statistics can provide an alternative to traditional artifact rejection based on arbitrary thresholds.

Here, we present a simple application of robust statistics to EEG data, from both a frequentist and Bayesian angle. The techniques we present here are well established in certain other research fields, but have not yet gained traction in the analysis of electrophysiological data. We demonstrate both the comparative ease of applying modern robust approaches and their ability to detect effects even in noisy EEG data.

Materials

The data used here were originally collected as part of a virtual reality experiment on the N400, with a multimodal semantic violation (Tromp et al. 2017). The experiment had two conditions, match and mismatch, with the mismatch eliciting the expected N400 in the original analysis.

The original data were recorded at 500 Hz and filtered online with a lowpass filter at 200 Hz and a highpass filter at 0.016 Hz (cf. Alday 2019 for a discussion of the online filter settings). The original study analyzed two N400 time windows (250–350ms and 350–600ms post stimulus onset) and found effects in both, albeit a weaker effect in the earlier time window. For the reanalysis reported here, we use the earlier time window to highlight that these techniques are not dependent on large effect sizes.

Methods

EEG Preprocessing

EEG data were preprocessed using MNE-Python v0.21 (Gramfort et al. 2013, 2014; Jas et al. 2017).

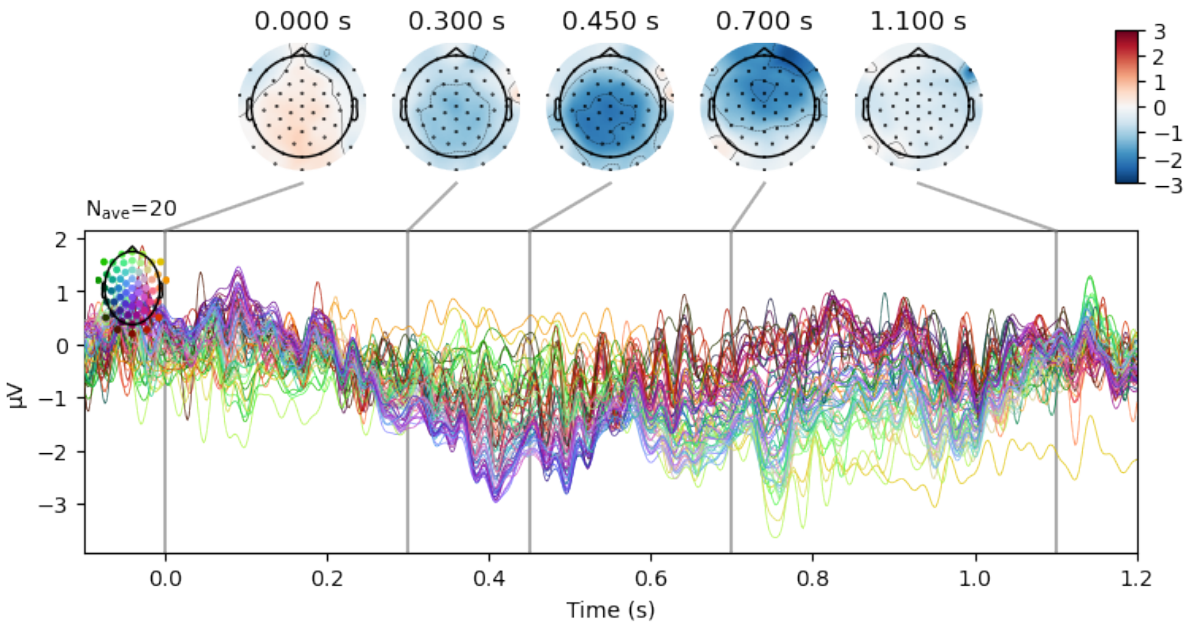

Figure 1: Difference wave for the “dirty” data without any artifact rejection

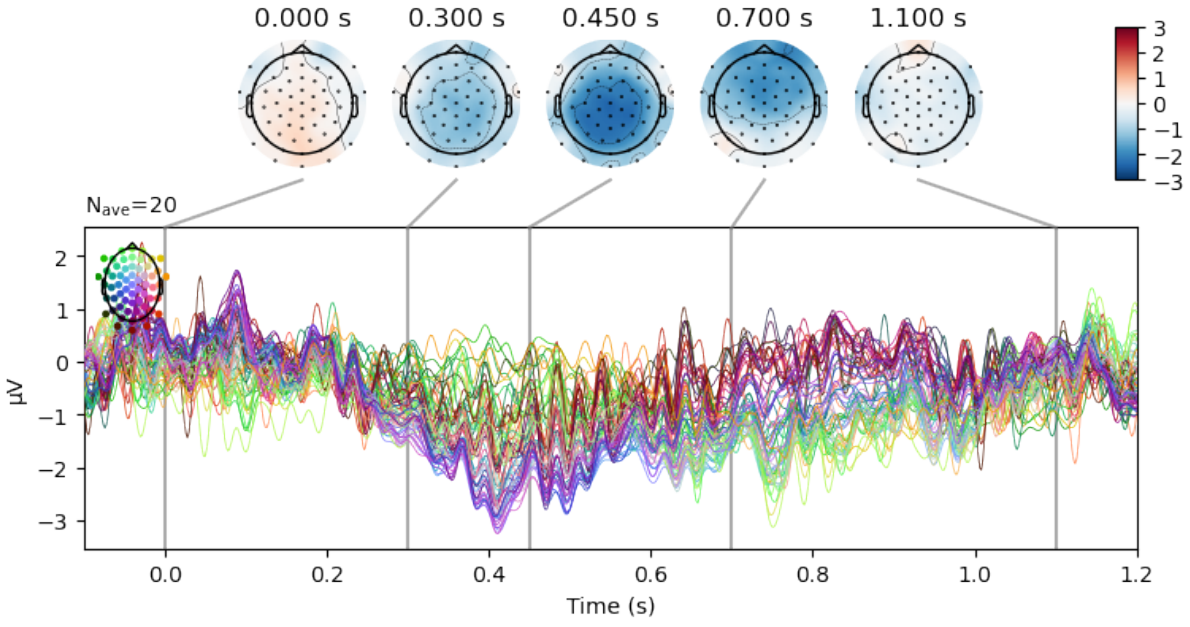

Figure 2: Difference wave for the “clean” data using peak-to-peak rejection with a 150µV threshold.

For simplicity of presentation, a stronger highpass filter (0.3 Hz) and no baseline correction were used. Traditional baseline correction has serious drawbacks and is often unnecessary. A better alternative is to use regression-based baseline correction, as demonstrated on this dataset in a previous re-analysis (Alday 2019). After applying a 0.3 Hz highpass filter however, whether or not baseline correction is performed does not change the pattern of results in this dataset. In other words, the tradeoff from using an appropriately designed highpass filter instead of a weaker filter with baseline correction is acceptable here (Alday 2019; Burkhard Maess, Schröger, and Widmann 2016; B. Maess, Schröger, and Widmann 2016; Widmann, Schröger, and Maess 2015).

Nevertheless, the supplementary materials contain an additional re-analysis of the data with regression-based baseline correction and a weaker (0.1 Hz) highpass filter. Similar results were observed, but the observed shifts (see Results) in model estimates occurred in the baseline predictor instead of the condition predictor.

For the "clean" data presented here, trials where any EEG or EOG channel exceeded 150µV peak-to-peak were excluded. This rejection threshold corresponds to the "default" threshold used in the MNE-Python documentation. For the "dirty" data, no artifact rejection of any kind was performed. This represents a worst-case scenario for robust methods.

Joint plots of the difference waves for the dirty and clean data are presented in Figures 1 and 2.

All analysis source code as well as the pre-processed single-trial data are available on OpenScience Framework. There are data for several different filter settings (0.1, 0.3, 0.5, 1.0 Hz highpass) as well as for several different baseline windows (500ms pre-stimulus, 200ms pre-stimulus, 100ms pre-stimulus, 200ms post-stimulus, average across entire epoch). Additionally, the maximum and minimum voltage in each epoch are reported, which allows for computing peak-to-peak and absolute-threshold based rejection offline.

For simplicity of presentation and to highlight the efficacy of robust methods, we used only the noisiest subset of the electrodes. We first computed the channel-wise standard error of the mean (for the mean amplitude in the 250–350ms time window) for each participant, then computed the root-mean-square error (RMSE) of that measurement (cf. "Standardized Measurement Error", Luck et al. 2020) across participants, resulting in a channel-level measurement of the noise in the ERP amplitude. Note that this provides an indication of the strength of the noise but not of the strength of the signal, and so even noisy channels may have a sufficient SNR for statistical analysis, especially when using robust methods. Only the 10 channels with a noise measurement above 1µV were used, as this corresponds to noise on the order of the average effect size in many language studies. These channels were all located in right anterior quadrant, which suggests that the high noise level may have been due to ocular artifacts and interference from the VR equipment. While this is not a typical N400 topography, the effects in the original study (see also Figures 1 and 2) were broadly distributed and symmetrical and are present in this quadrant. Finally, by focusing on only these three electrodes, we can omit topographical analyses, which are comparatively uninteresting for a well-established effect such as the N400. Moreover, there is no single consensus as to the ideal topographical analysis (cf. Kretzschmar and Alday, n.d.) and addressing that problem is beyond the scope of the present work.

After filtering and, for the clean data, artifact rejection, the single-trial amplitude between 250 and 350ms post-stimulus (the early time window used in the original analysis) was averaged across these electrodes.

Statistical Analysis

For all statistical analyses, we used mixed-effects models of single-trial data. For language experiments, it is critical to account for both subject and item effects (Coleman 1964; Clark 1973; Baayen, Davidson, and Bates 2008; Judd, Westfall, and Kenny 2012). The behavioral side of the field has recognized this and started to incorporate random effects into their analyses, but ERP researchers have been slow to catch up (cf. Bürki, Frossard, and Renaud 2018). However, modeling subject and item effects is just as essential in ERP analyses as it is in behavioral paradigms. Mixed-effects models provide a natural solution to this problem.

Fixed effects consisted only of condition, which was contrast coded using effects (±1) coding. The reference level was "match", so that the coefficient estimate reflects the direction of the change in amplitude (i.e. a negativity). Random effects consisted of by-participant intercepts, by-item intercepts and by-participant slopes for condition. This is equivalent to the assumption that participants may differ both in their overall EEG response and the strength of their response to the mismatch and that items may differ in the strength of the

response they elicit. Given the coarse spatial granularity of ROIs, we do not expect any between-participant or between-item differences and omit ROI from the random effects for parsimony. Similarly, preliminary analyses suggested that by-item slopes for condition leads to an overparameterized model and we omit them from the model for parsimony.

The R programming analysis was used for all analyses (version 4.0.3).

Frequentist Analysis

We used the R packages `lme4` (version 1.1-23, Bates et al. 2015) and `robustlmm` (version 2.3, Koller 2016) for the frequentist analysis. For the classical, non-robust analyses, models were fit with maximum-likelihood estimation. For the robust analysis, robustness was achieved by robustification of the scoring equations (i.e. the gradient of the log-likelihood) (Koller 2016). Conceptually, a classical, non-robust model weighs the residual error for each observation quadratically (i.e. observations that lie further from the model's prediction have an outsize influence on the coefficient estimates, "pulling" the model predictions in their direction) whereas the robust model weighs residual error for inliers quadratically, but residual error for outliers linearly (i.e. if an observation is classified as an outlier, it does not "pull" as hard on the model predictions). In order to determine which observations are inliers and which are outliers, the model is fit multiple times in an iterative process, each time improving its classification of in- versus outliers. The major downside to this approach is that the fitting algorithm no longer corresponds to any likelihood and as such model comparison – whether through likelihood-ratio tests or information criteria such as AIC – is no longer straightforward, as the usual basis for comparison is no longer defined.

Bayesian Analysis

For the Bayesian analysis, we used `brms` (version 2.14, Bürkner 2017, 2018). For the standard analysis, a Gaussian (normal) likelihood was used. The *likelihood* here corresponds to the distribution of the residual error term, which is generally called the model `family` in various statistical software packages. A Gaussian likelihood corresponds to classical linear regression, where the residuals are assumed to be normally distributed. For the robust analysis, a Student-t likelihood was used. The t distribution is similar to the normal distribution, but with heavier tails and as such is used in both Bayesian and non-Bayesian modeling for robust statistics (cf. the overview in Koller 2016). In abstract terms, the t-distribution can be viewed as a mixture of normal distributions with differing variance (for our purposes, a mixture of "inlier" and "outlier" distributions). In practical terms, the t-distribution is like a normal distribution where the proportion of extreme values (i.e. outliers) is higher. The underlying mixture, or equivalently, the proportion of outliers is reflected in the degrees of freedom, nu, with higher values of nu corresponding to fewer outliers. In the limiting case as the Student-t degrees of freedom goes to infinity, the proportion of outliers goes to zero and the t-distribution becomes a normal distribution. This property can also be observed at around nu=30, which is the origin of many rules-of-thumb taught in introductory statistics courses. In the other limiting case, as nu goes towards zero, the t-distribution becomes a Cauchy distribution, which has no well-defined mean. This corresponds roughly to the "all outliers" or exceptionally noisy data case, where it is not possible to make any inference about the mean (and hence about the effect).

For all Bayesian models, default priors were used, which we present in overview here. For more details (e.g. of the priors on the residual variance), we refer the interested reader to the brms documentation. In particular, flat priors were used for fixed effects. Generally, the use of default, flat priors is not preferred in Bayesian analyses; however, this provides the most straightforward comparison to classical analyses using frequentist techniques. For the random effects, half Student $t(3)$ priors for the standard deviation of the random effects (a *very* weakly regularizing prior), LKJ(1) priors for the correlation of random effects (corresponding to equal probability of valid correlations, i.e. values on [-1,1]). For nu, the degrees of freedom on the Student-t likelihood in the robust model, a $\Gamma(2, 0.1)$ prior with a fixed lower bound of 1 was used, corresponding to a preference for smaller values.

After sampling, all chains had $\hat{R} = 1$ and both bulk and tail effective sample sizes greater than 3000 samples for all parameters.

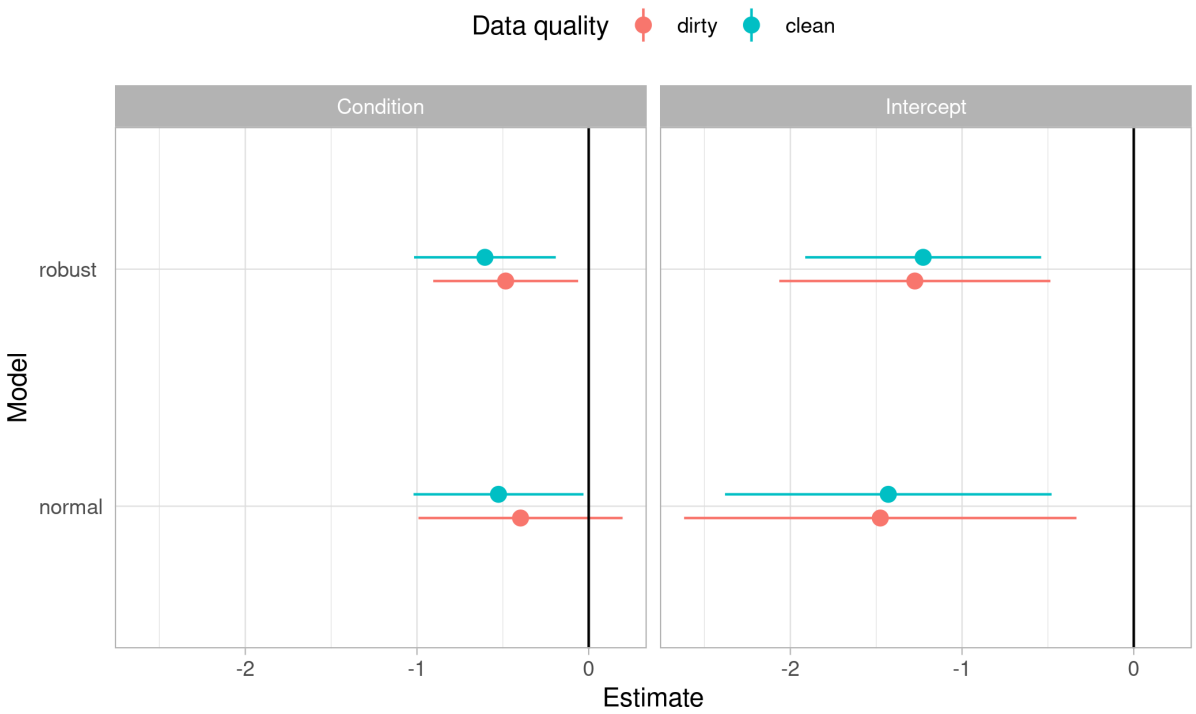

Figure 3: Estimates and Wald 95% confidence intervals for the frequentist models. Confidence intervals were computed as the estimate plus or minus twice the standard error. No estimates are provided for the residual error, because there is no efficient and accurate way to compute the confidence interval on that estimate for robust models. All models provide similar estimates, but the robust model provides narrower confidence intervals for both clean an dirty data. In other words, the robust model is a more powerful procedure for imperfect data.

Results

For all models (robust and non-robust, frequentist and Bayesian) and data (clean and dirty), we present the point estimates and uncertainty estimates (intervals for the frequentist models, posterior distributions for the Bayesian models). Additionally, the summary output (generated with `summary` in R) is shown for comparison.

Frequentist Analysis

The results of the frequentist analysis can be found in Figure 3 as well as Tables 1 and 2.

Table 1: Classical (non-robust) frequentist summaries. The estimate of the residual standard deviation as well as the standard error on condition differ between the clean and dirty data. Non robust methods are sensitive to the outlier-like nature of artifacts. Note that the standard error of the estimates for the random effects are generally not reported for frequentist models, as the sampling distribution is known to be highly skewed.

	Cleaned data		Dirty data	
	Estimate	SE	Estimate	SE
Intercept by Item (standard deviation scale)	.98		1.52	
Intercept by Participant (standard deviation scale)	1.87		2.24	
Condition by Participant (standard deviation scale)	0.66		0.91	
Correlation(Intercept, Condition) by Participant	-0.25		-0.38	

	Cleaned data		Dirty data	
	Estimate	SE	Estimate	SE
Intercept	-1.43	0.48	-1.48	0.57
Condition	-0.53	0.25	-0.40	0.30
Correlation(Intercept, Condition)	-0.13		-0.23	
Residual standard deviation	7.59		8.65	
AIC	10124.0		11502.6	
BIC	10161.0		11540.2	
Log-likelihood	-5055.0		-5744.3	
Deviance	10110.0		11488.6	

Table 2: Robust frequentist model summaries. The estimate of the residual standard deviation as well as the standard error on condition do **not** differ between the clean and dirty data. Robust methods are less sensitive to the outlier-like nature of artifacts. In the dirty data, the number of downweighted observations has increased. No likelihood-based descriptions of the fit (log likelihood, deviance, AIC, BIC) are reported as the robust fit does not correspond to any likelihood or pseudo-likelihood (Koller 2016). Note that the standard error of the estimates for the random effects are generally not reported for frequentist models, as the sampling distribution is known to be highly skewed.

	Cleaned data		Dirty data	
	Estimate	SE	Estimate	SE
Intercept by Item (standard deviation scale)	0.00		0.00	
Intercept by Participant (standard deviation scale)	1.21		1.48	
Condition by Participant (standard deviation scale)	0.23		0.32	
Correlation(Intercept, Condition) by Participant	-1.00		-1.00	
Intercept	-1.23	0.34	-1.27	0.39
Condition	-0.60	0.21	-0.48	0.21
Correlation(Intercept, Condition)	-0.20		-0.31	
Residual standard deviation	7.43		7.68	
Number of downweighted residuals (robustness weights not equal to 1)	309		331	
Number of downweighted random effects (robustness weights not equal to 1)	4		4	
Rho functions:				
Rho functions, residuals: smoothed Huber k, s	1.345, 10		1.345, 10	
Rho functions, by item: smoothed Huber k, s	1.345, 10		1.345, 10	
Rho functions, by participant: smoothed Huber k, s	1.345, 10		1.345, 10	

The model summaries for robust models are similar in information to the summaries for models fitted with maximum-likelihood estimation, with a few key differences: First, there are no likelihood-based statistics (i.e. AIC, BIC, log-likelihood, deviance). Second, the the model summary includes information about rho functions, which provide insight into the robustness procedure. The smoother Huber functions essentially convert the measure of central tendency from the mean to a mixture of the mean and the median; or equiva-

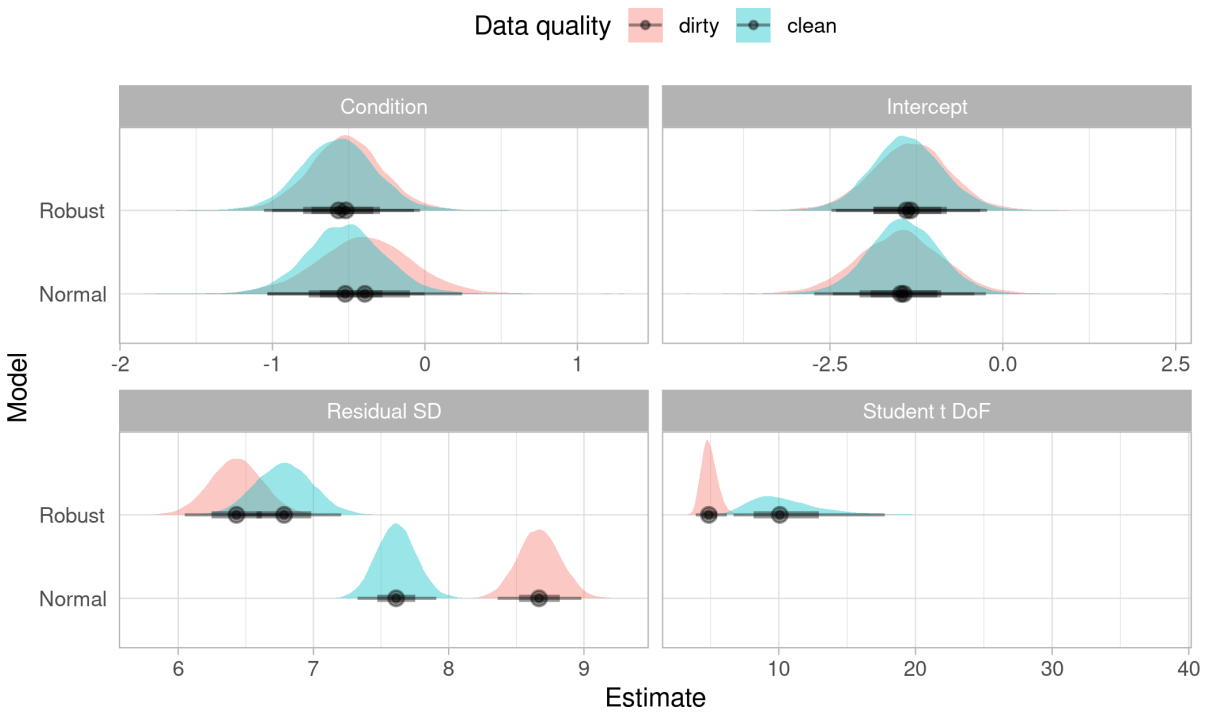

Figure 4: Posterior distributions from the Bayesian models. The condition estimates do not vary much between the clean and the dirty data for the robust model, but do for the normal model. The estimates for residual error and the the Student-t degrees of freedom differ between both between datasets and between models

lently downweighting extreme values. The k parameter determines the threshold for transitioning from the mean to the median (i.e. transitioning from inlier to outlier), while the s parameter controls the smoothness of that transition. As k increases, the efficiency (in the technical sense) of the estimates increases, but the robust decreases, with $k \to \infty$ being equivalent to REML estimation. The values here are the default values ($k = 1.345$, $s = 10$) and corresponds to 95% efficiency of the non-robust procedure. Between the classic model summary and the description of the rho functions, the robustness weights describes what number of reweighted observations and random-effects levels. We note that even the clean data has a non trivial proportion of data with robustness weights not equal to one – some observations are always more influential than others, and robust regression decreases the amount of excess influence.

Notably, the maximum-likelihood and robust methods provided similar estimates for the clean data, but the robust method provided smaller standard errors. The random effect for item was shrunk to zero by the robust method, but not the non-robust method. This does not imply that the by-item variation was zero, but rather that the by-item variation in the observed data (here: right anterior electrodes) is not distinguishable from the observation-level (residual) variation. In other words, for the clean data, the robust technique was as effective as the non-robust technique.

For the dirty data, the estimates between methods were again similar, but not identical. The standard errors for the robust model were nearly a third smaller, with the result that the t-value was larger. Using $|t| > 2$ as an asymptotic approximation to the traditional 5% significance level (cf. Baayen, Davidson, and Bates 2008), the robust model detects the effect, but the non-robust model does not. In other words, the robust model has *more statistical power* when the data are not cleaned.

Bayesian Analysis

The results of the Bayesian analysis can be found in Figure 4 as well as Tables 3 and 4.

Table 3: Gaussian (non-robust) Bayesian model summaries. The estimate of the residual standard deviation as well as the standard error on condition differ between the clean and dirty data. Non robust methods are sensitive to the outlier-like nature of artifacts.

	Cleaned data		Dirty data	
	Estimate	SE	Estimate	SE
Intercept by Item (standard deviation scale)	0.87	0.39	1.50	0.36
Intercept by Participant (standard deviation scale)	2.09	0.44	2.50	0.51
Condition by Participant (standard deviation scale)	0.65	0.35	0.98	0.39
Correlation(Intercept, Condition) by Participant	-0.19	0.40	-0.31	0.32
Intercept	-1.43	0.52	-1.48	0.63
Condition	-0.52	0.26	-0.39	0.32
Residual standard deviation (sigma)	7.61	0.15	8.67	0.16

Table 4: Student-t (robust) Bayesian model summaries. The estimate of the residual standard deviation as well as the standard error on condition do **not** differ between the clean and dirty data. Robust methods are less sensitive to the outlier-like nature of artifacts. In the dirty data, the estimate for the nu parameter has decreased, corresponding to a larger number of observations being treated as outlier-like.

	Cleaned data		Dirty data	
	Estimate	SE	Estimate	SE
Intercept by Item (standard deviation scale)	0.97	0.37	1.52	0.29
Intercept by Participant (standard deviation scale)	2.11	0.44	2.25	0.46
Condition by Participant (standard deviation scale)	0.60	0.35	0.60	0.34
Correlation(Intercept, Condition) by Participant	-0.07	0.43	-0.06	0.41
Intercept	-1.38	0.53	-1.34	0.57
Condition	-0.57	0.25	-0.52	0.25
Residual standard deviation (sigma)	6.78	0.21	6.43	0.20
Student-t degrees of freedom (nu)	10.62	2.88	4.93	0.59

The primary difference between the summaries of the Gaussian and Student-t likelihoods is the presence of the nu parameter. As mentioned above, the nu parameter is the degrees of freedom in the Student-t distribution, with a smaller value of nu corresponding to a more heavy-tailed distribution. A value of nu above approximately 30 corresponds to a distribution which is essentially indistinguishable from the Gaussian. We note that the estimate for nu for both the clean and the dirty data is far smaller than 30, suggesting, as for the frequentist results, that some observations may be particularly influential in a classical analysis.

For the clean data, the two models provided estimates identical to the first decimal point. The credible interval for the Student-t likelihood clearly does not cross zero, while the credible interval for the Gaussian likelihood just about reaches zero. (Note that the upper edge is estimated at negative zero, which indicates here an estimate infinitesimally smaller than, but functionally equivalent to, zero.) In a full Bayesian analysis, it is generally preferable to define a region of practical equivalence (ROPE, Kruschke and Liddell 2017; Kruschke 2018) rather than compare a credible interval to a point hypothesis. Nonetheless, from a Bayesian estimation perspective (cf. Kruschke and Liddell 2017), the narrower credible interval suggests that the Student-t model provides a more precise estimate . In other words, for the clean data, the robust technique was just as, if not more, effective as the non-robust technique.

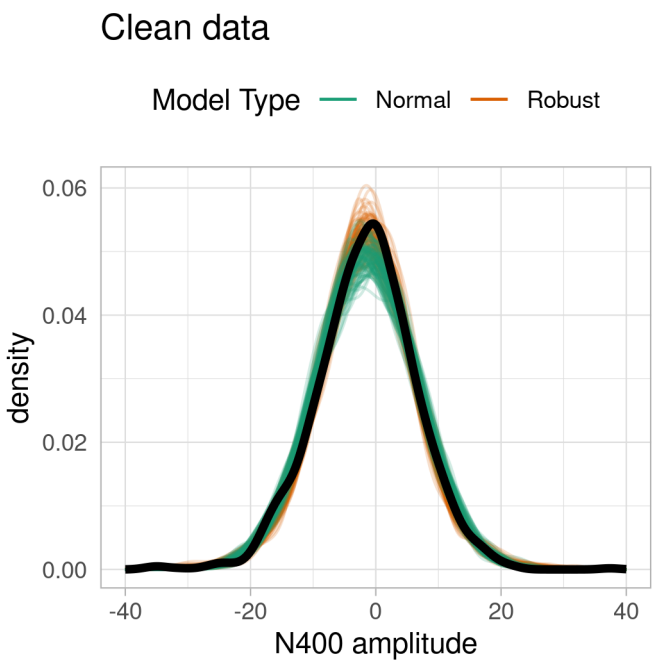

Figure 5: Posterior predictive distribution for the Bayesian models and the clean data. Both models capture the overall pattern of the data equally well.

For the dirty data, the two models again provide similar estimates, albeit less similar than for the clean data. The credible interval is slightly narrower for the Student-t model, resulting in a credible interval for the condition effect that does not cross zero. In other words, the robust model provides higher precision and thus more power when the data have not been cleaned.

Figures 5 and 6 present posterior predictive checks for the clean and dirty data, respectively. Both models capture the overall pattern of the clean data well, but the Student-t likelihood performs better for the dirty data. Because the Gaussian likelihood is unable to accommodate outliers without increasing its variance, it must necessarily shift density away from the mean, resulting in a poorer fit for inliers.

Discussion

In the analyses presented here, we demonstrated how modern robust techniques can perform as well as traditional techniques for both clean and dirty data. In the following, we briefly discuss some key implications and recommendations for using robust techniques in ERP research.

Robust statistics: a good default, but not a panacea

The data presented above were selected to be noisy (via choice of channels in an experiment with virtual reality) and have a less-than-maximal effect (via choice of time window). We elected to use suboptimal data in order to illustrate the benefits of using robust statistics, but having low signal to noise ratio in EEG data is not at all an unrealistic scenario. Neuroscience studies are chronically underpowered and EEG in general and ERP in particular is susceptible to many sources of noise (cf. Boudewyn et al. 2017; Clayson et al. 2019; Button et al. 2013). Even when using a standard artifact rejection technique, the robust techniques performed on par with the traditional techniques. Both of the robust techniques (frequentist as well as Bayesian) used here are asymptotically equivalent to the traditional technique. As such, they will perform exactly the same when there is sufficient, well-behaved data. Crucially, however, we generally do not know in advance if we have sufficient, well-behaved data. This suggests that there is little to be lost in using robust techniques, but potentially much to be gained. As such, robust techniques present a reasonable default for most ERP analyses.

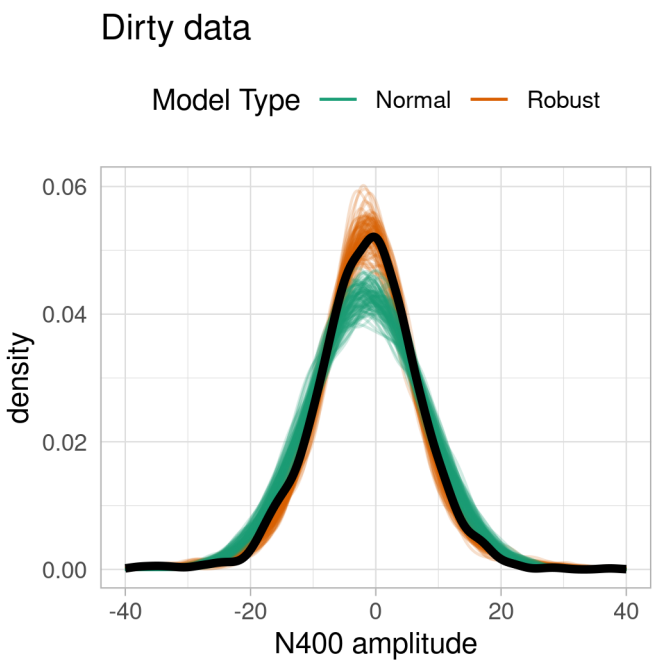

Figure 6: Posterior predictive distribution for the Bayesian models and the dirty data. The traditional Gaussian likelihood overestimates the tails and underestimates the mode, reflecting its sensitivity to outliers. The Student-t likelihood is able to accomodate outliers without moving density away from the mode, resulting in a better match to the overall pattern of the data.

Despite the advantages we describe, robust statistics are not a panacea. While robust statistics can handle outliers and their negative impact on statistical power and inference better than non-robust methods, they are not a replacement for designing for sufficient statistical power in the first place. Moreover, it is far more efficient to prevent outliers from occurring than to deal with them after the fact. In other words, the best inferences are built upon a large amount of high quality data. The best way to have a large amount of high quality data available is collect a large amount of data carefully, by constructing experiments with minimal confounds, reducing environmental noise, using a setup that discourages EOG and EMG artifacts, etc.

Robust statistics *and* artifact detection: a winning combination

Building upon this perspective, we recommend robust statistics as a complement to traditional taxonomic approaches to artifacts and not, in general, as a replacement for them. For example, eye-movements can often be very well characterized and removing or correcting them (Gratton, Coles, and Donchin 1983; Jung et al. 1997) can be an effective technique, decreasing the portion of the EEG signal not arising from brain activity. Similarly, segments clearly reflecting a bad electrode connection, ("CRAP" in the terminology of Luck 2005) can also be safely removed as they will generally contain little to signal and greatly add to the aggregate noise. Likewise, signals reflecting biologically implausible signals (e.g. amplitudes of several hundred microvolts) can also be excluded.[1] However, more arbitrary thresholds such as absolute and peak-to-peak voltage or slope can be set to more permissive values, thus not drawing as sharp a line between extreme signal value and true artifactual value, if robust statistics are used. In other words, robust models are capable of extracting what information is available from more extreme values without breaking down. Robust statistics can then capture the full variation in the data and not just the variation in a narrow band of the data assumed to be representative by a researcher. By setting looser thresholds, whether algorithmic or manually annotated, the data analysis procedure becomes less dependent on a particular researcher and researcher degrees of freedom, reducing researcher degrees of freedom and improving the reproducibility of ERP analyses (Gelman and Loken 2013; Simmons, Nelson, and Simonsohn 2011)

Robust statistics thus enable analyzing comparatively noisy data, such as from experiments with populations

[1] We note that this is actually a Bayesian procedure that could be implemented quantitatively and algorithmically in a Bayesian robust model.

where motion and motor artifacts are common. Studies on infants and clinical populations are particularly rich application areas for robust statistics because data collection is difficult (thus sample sizes are often small, rendering violations of assumptions particularly problematic), inter-individual variability often high (thus violating homoskedacity assumptions in non-robust methods), and the EEG noisy (e.g. from motor artifacts related to trembling in individuals with Parkinson's disease). Robust statistics provide a promising complement to traditional taxonomic approaches to artifact correction and are especially useful when analyzing data from new types of experiments and historically understudied populations.

A word of caution on comparing model fits

In our comparison of robust and traditional techniques, we stated that some model comparison techniques such as the classical likelihood ratio test and information criteria based on maximum-likelihood estimate (e.g. AIC, BIC) are not valid for robust models. When presenting our reanalysis, we did not compare the robust and traditional fits directly, instead comparing only their estimates and, for the Bayesian models, visually inspecting the posterior-predictive checks. This is intentional. Measures such as the coefficient of determination (R^2) are not trivial to define for mixed-effects models (at least not so that they maintain all of the properties they have for classical fixed-effects regression). Moreover, the definitions of such measures are based on means and therefore tend to be biased towards traditional, non-robust methods, even when those methods provide a worse summary of the data. In particular, classical ordinary least squares (OLS) regression is the best linear unbiased estimator (BLUE) via the Gauss-Markov theorem and therefore will be better at maximizing the R^2 than any other technique. However desirable that may sound, BLUE is not always the optimal estimator for a given task. (For example, Stein (1956) showed that biased estimators will outperform unbiased estimators in predicting out-of-sample data.) In more concrete terms, the downweighting of outliers in robust methods reduces the fit to the observed data because the estimates are moved further from the outliers, but this bias will tend produce better performance with future data (i.e. robust methods result in better out-of-sample prediction). Instead of depending on measures defined to favor the traditional techniques, we instead recommend examining how well the model captures patterns in the data, via e.g. posterior-predictive checks and plotting fitted vs. observed data. George Box's famous aphorism states that all models are wrong, but some models are useful. The most useful model is one that not only captures patterns in the observed data, but also generalizes well to not-yet observed data, even if that means being *slightly* worse at describing the observed data.

Conclusion

Traditional approaches to ERP analysis make a strong distinction between artifact detection and statistical analysis with the ostensible goal of improving the quality of the statistical analysis. Here, we demonstrated that an integrative approach using robust statistics can improve the quality of the analysis even for artifact free data. Moreover, robust statistics reduce the need for arbitrary thresholds, while enabling the analysis of noisy data from experiments outside the traditional laboratory environment and beyond the traditional population of healthy, young adult university students. Non-robust methods have long been the unquestioned default for analyzing ERP experiments, but given the ease of use and potentially improved statistical power of robust methods, that default seems hard to defend. We recommend researchers at the very least explore and report robust models in addition to (if not in place of) classical, non-robust models.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

PA and JvP conceptualized the study. PA designed and performed the analysis. PA and JvP wrote the paper.

References

Alday, Phillip M. 2019. “How Much Baseline Correction Do We Need in ERP Research? Extended GLM Model Can Replace Baseline Correction While Lifting Its Limits.” *Psychophysiology* 56 (12). https://doi.org/10.1111/psyp.13451.

Baayen, R. H., D. J. Davidson, and D. M. Bates. 2008. “Mixed-Effects Modeling with Crossed Random Effects for Subjects and Items.” *Journal of Memory and Language* 59: 390–412.

Bates, Douglas, Martin Maechler, Benjamin M. Bolker, and Steven Walker. 2015. “Fitting Linear Mixed-Effects Models Using Lme4.” *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Boudewyn, Megan A., Steven J. Luck, Jaclyn L. Farrens, and Emily S. Kappenman. 2017. “How Many Trials Does It Take to Get a Significant Erp Effect? It Depends.” *Psychophysiology*, e13049. https://doi.org/10.1111/psyp.13049.

Bürki, Audrey, Jaromil Frossard, and Olivier Renaud. 2018. “Accounting for Stimulus and Participant Effects in Event-Related Potential Analyses to Increase the Replicability of Studies.” *Journal of Neuroscience Methods*, September. https://doi.org/10.1016/j.jneumeth.2018.09.016.

Bürkner, Paul-Christian. 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.

———. 2018. “Advanced Bayesian Multilevel Modeling with the R Package brms.” *The R Journal* 10 (1): 395–411. https://doi.org/10.32614/RJ-2018-017.

Button, Katherine S, John P A Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma S J Robinson, and Marcus R Munafò. 2013. “Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience.” *Nat Rev Neurosci*. https://doi.org/10.1038/nrn3475.

Clark, Herbert H. 1973. “The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research.” *Journal of Verbal Learning and Verbal Behavior* 12: 335–59. https://doi.org/10.1016/S0022-5371(73)80014-3.

Clayson, Peter E., Kaylie A. Carbine, Scott A. Baldwin, and Michael J. Larson. 2019. “Methodological Reporting Behavior, Sample Sizes, and Statistical Power in Studies of Event-Related Potentials: Barriers to Reproducibility and Replicability.” *Psychophysiology*, July. https://doi.org/10.1111/psyp.13437.

Coleman, E. B. 1964. “Generalizing to a Language Population.” *Psychological Reports* 14 (1): 219–26. https://doi.org/10.2466/pr0.1964.14.1.219.

Dimigen, Olaf. 2020. “Optimizing the ICA-Based Removal of Ocular EEG Artifacts from Free Viewing Experiments.” *NeuroImage* 207 (February): 116117. https://doi.org/10.1016/j.neuroimage.2019.116117.

Gelman, Andrew, and Eric Loken. 2013. “The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No ‘Fishing Expedition’ or ‘P-Hacking’ and the Research Hypothesis Was Posited Ahead of Time.” http://www.stat.columbia.edu~gelman/research/unpublished/p_hacking.pdf.

Gramfort, Alexandre, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, et al. 2013. “MEG and Eeg Data Analysis with Mne-Python.” *Front Neurosci* 7: 267. https://doi.org/10.3389/fnins.2013.00267.

Gramfort, Alexandre, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Lauri Parkkonen, and Matti S. Hämäläinen. 2014. “MNE Software for Processing Meg and Eeg Data.” *NeuroImage* 86: 446–60. https://doi.org/10.1016/j.neuroimage.2013.10.027.

Gratton, Gabriele, Michael G.H Coles, and Emanuel Donchin. 1983. “A New Method for Off-Line Removal of Ocular Artifact.” *Electroencephalography and Clinical Neurophysiology* 55 (4): 468–84. https://doi.org/https://doi.org/10.1016/0013-4694(83)90135-9.

Jas, Mainak, Eric Larson, Denis-Alexander Engemann, Jaakko Leppakangas, Samu Taulu, Matti Hamalainen, and Alexandre Gramfort. 2017. “MEG/Eeg Group Study with Mne: Recommendations, Quality Assessments and Best Practices.” https://doi.org/10.1101/240044.

Judd, Charles M., Jacob Westfall, and David A. Kenny. 2012. “Treating Stimuli as a Random Factor in Social Psychology: A New and Comprehensive Solution to a Pervasive but Largely Ignored Problem.” *J Pers Soc Psychol* 103 (1): 54–69. https://doi.org/10.1037/a0028347.

Jung, Tzyy-Ping, Colin Humphries, Te-Won Lee, Scott Makeig, Martin J. McKeown, Vicente Iragui, and Terrence J. Sejnowski. 1997. "Extended Ica Removes Artifacts from Electroencephalographic Recordings." *Advances in Neural Information Processing Systems* 10.

Koller, Manuel. 2016. "robustlmm: An R Package for Robust Estimation of Linear Mixed-Effects Models." *Journal of Statistical Software* 75 (6): 1–24. https://doi.org/10.18637/jss.v075.i06.

Kretzschmar, Franziska, and Phillip M. Alday. n.d. "Principles of Statistical Analysis: Old and New Tools." In *Language Electrified. Techniques, Methods, Applications, and Future Perspectives in the Neurophysiological Investigation of Language*, edited by Mirko Grimaldi, Yury Shtyrov, and Elvira Brattico. https://doi.org/10.31234/osf.io/nyj3k.

Kruschke, John K. 2018. "Rejecting or Accepting Parameter Values in Bayesian Estimation." *Advances in Methods and Practices in Psychological Science* 1 (2): 270–80. https://doi.org/10.1177/2515245918771304.

Kruschke, John K., and Torrin M. Liddell. 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.

Luck, Steven J. 2005. *An Introduction to the Event-Related Potential Technique*. Cambridge, MA: MIT Press.

Luck, Steven J, Andrew X Stewart, Aaron Matthew Simmons, and Mijke Rhemtulla. 2020. "Standardized Measurement Error: A Universal Measure of Data Quality for Averaged Event-Related Potentials (V25)," May. https://doi.org/10.31234/osf.io/dwm64.

Maess, B., E. Schröger, and A. Widmann. 2016. "High-Pass Filters and Baseline Correction in M/Eeg Analysis–Continued Discussion." *Journal of Neuroscience Methods*, –. https://doi.org/10.1016/j.jneumeth.2016.01.016.

Maess, Burkhard, Erich Schröger, and Andreas Widmann. 2016. "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 (June): 164–65. https://doi.org/10.1016/j.jneumeth.2015.12.003.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology : Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66. https://doi.org/10.1177/0956797611417632.

Stein, Charles. 1956. "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 197–206. Berkeley, Calif.: University of California Press. http://projecteuclid.org/euclid.bsmsp/1200501656.

Tromp, Johanne, David Peeters, Antje S. Meyer, and Peter Hagoort. 2017. "The Combined Use of Virtual Reality and Eeg to Study Language Processing in Naturalistic Environments." *Behavior Research Methods*. https://doi.org/10.3758/s13428-017-0911-9.

Widmann, Andreas, Erich Schröger, and Burkhard Maess. 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.

Wilcox, Rand R. 2010. *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*. 2. ed. New York: Springer. https://doi.org/10.1007/978-1-4419-5525-8.

———. 2012. *Introduction to Robust Estimation and Hypothesis Testing*. 3. ed. Academic Press.