Title: Accounting for stimulus and participant effects in Event-related potential analyses to increase the replicability of studies

Authors: Audrey Bürki, Jaromil Frossard, Olivier Renaud

PII: S0165-0270(18)30277-2
DOI: https://doi.org/10.1016/j.jneumeth.2018.09.016
Reference: NSM 8118

To appear in: *Journal of Neuroscience Methods*

Received date: 28-2-2018
Revised date: 21-8-2018
Accepted date: 6-9-2018

Please cite this article as: Bürki A, Frossard J, Renaud O, Accounting for stimulus and participant effects in Event-related potential analyses to increase the replicability of studies, *Journal of Neuroscience Methods* (2018), https://doi.org/10.1016/j.jneumeth.2018.09.016

Accounting for stimulus and participant effects in Event-related potential analyses to increase the replicability of studies

Running head: **Stimulus and participant as random effects in ERP analyses**

Audrey Bürki [a, b*], Jaromil Frossard [a], & Olivier Renaud [a]

[a] Methodology and Data analysis, Section of Psychology, FPSE, University of Geneva, Bd du Pont d'Arve 42, 1205 Genève, Switzerland

[b] Cognitive Sciences, Department of Linguistics, University of Potsdam, Karl-Liebknecht-Straße 24-25, 14476 Potsdam, Germany, buerki@uni-potsdam.de

*corresponding author

Highlights

- Paper focusses on statistical analysis of ERPs in cognitive Neuroscience
- Current practices do not account for stimulus as random effect
- Simulations show high Type I error rates of such analyses
- Alternative methods are proposed, their validity is demonstrated
- Failure to change current practices fosters the replication crisis

Abstract

Background

Event-related potentials (ERPs) are increasingly used in cognitive science. With their high temporal resolution, they offer a unique window into cognitive processes and their time course. In this paper, we focus on ERP experiments whose designs involve selecting participants and stimuli amongst many. Recently, Westfall, Nichols, and Yarkoni (2017) highlighted the drastic consequences of not considering stimuli as a random variable in fMRI studies with such designs. Most ERP studies in cognitive psychology suffer from the same drawback.

New Method

We advocate the use of the Quasi-F or Mixed-effects models instead of the classical ANOVA/by-participant F1 statistic to analyze ERP datasets in which the dependent variable is reduced to one measure per trial (e.g., mean amplitude). We combine Quasi-F statistic and cluster mass tests to analyze datasets with multiple measures per trial. Doing so allows us to treat stimulus as a random variable while correcting for multiple comparisons.

Results

Simulations show that the use of Quasi-F statistics with cluster mass tests allows maintaining the family wise error rates close to the nominal alpha level of 0.05.

Comparison with existing methods

Simulations reveal that the classical ANOVA/F1 approach has an alarming FWER, demonstrating the superiority of models that treat both participant and stimulus as random variables, like the Quasi-F approach.

Conclusions

Our simulations question the validity of studies in which stimulus is not treated as a random variable. Failure to change the current standards feeds the replicability crisis.

Keywords: Cluster Mass; ERP; quasi-F; mixed-effects model; replicability crisis; Stimulus as fixed-effect fallacy

1. Introduction

Many subfields in cognitive psychology consider Event-Related Potentials (ERPs) as a measurement of choice to study cognitive processes. ERPs are particularly enticing because of their high temporal resolution, which offers a unique window into the time course and nature of cognitive processes. ERP data come with methodological challenges, however. They are complex, and their analysis requires specific statistical models. Because they are often analyzed using inadequate models (see below), it can be expected that the replicability rate of these studies is low, probably much lower than the replicability rate of psychological experiments in which the dependent measure is much simpler (i.e., one measure per trial as opposed to many in an ERP experiment) and this rate is nothing to be proud of (e.g., Simmons, Nelson, & Simonsohn, 2011).

The aim of this paper is to review and question existing practices in the analysis of ERP data, show the potential pitfalls of current standards, review better suited statistical tools, and discuss some of the current challenges in the statistical analysis of such data. The paper complements recent contributions on various aspects of the statistical analysis of ERP data (see for instance Keil et al., 2014; Luck & Gaspelin, 2017) by focusing on the need to take into account the fact that the stimuli of the experiment have been selected amongst many available options (e.g., subset of pictures, words, animal sounds), or, in more technical terms, to treat stimulus as a random effect.

The need to treat stimulus as a random effect was highlighted early on in the literature (e.g., Clark, 1973) and is since then regularly discussed (e.g., Judd, Westfall, & Kenny, 2012). Recently, Westfall, Nichols, and Yarkoni (2017) reported simulations suggesting that many of the results reported as

significant in fMRI studies likely result from Type I errors (i.e., rejecting the null hypothesis when it is in fact true), as a direct consequence of not including stimulus as a random effect in statistical models. In this contribution, we examine this issue in the context of ERP studies. Treating stimulus as a random effect in the context of ERP studies raises specific issues. Moreover, given the many forms that an ERP data set can take, there is not one but many statistical approaches. In this paper we review the options available for different types of data sets.

The paper is structured as follows. In the next section, we describe the properties of data sets collected in typical psychological experiments, in which participants and stimuli are selected amongst many. We discuss the options that are used and/or available to the researcher to take into account these properties in the statistical analysis when the dependent measure consists in one measurement per trial (e.g., response type, response time, viewing time). In section 3, we extend the discussion to ERP experiments. We discuss the statistical tools used or available to analyze data sets involving one to many measures in the time dimension. We demonstrate, using simulations, the consequences of not taking the experiment design into account. In the last section, we discuss current challenges in the extension of existing tools to data sets with several dozens/hundreds of measurements in both the time and space dimensions.

2. Designs in which participants and stimuli are selected amongst many

In most experiments in cognitive psychology, the researcher is interested in the effect of one or several independent variables (or predictors, also called fixed-effects in the context of the statistical model) on a dependent measure. Examples of dependent measures are response times (time interval between the onset of the stimulus presentation and the onset of the response), time spent gazing at a given location, probability of giving a correct response, or of remembering an item. Independent variables may code for properties of the design (e.g., whether the stimulus is preceded by a related or unrelated word, number of repetitions of a given stimulus, specific condition in which the person is set, group category, etc.) or of the stimuli (e.g., sentence type in language-

related experiments, emotion associated with a given picture, typicality or complexity of the picture, etc.).

In many cases, and the present paper is specifically concerned with those cases, researchers build up experiments that include sets of stimuli (these can be words, pictures, shapes, situations, sounds, etc.) and sets of participants. Stimuli and participants are selected amongst many alternatives. For instance, the experiment will typically involve two or three dozens of participants, selected in a larger pool of psychology students. The researcher is usually not interested in this subset of participants per se, but would like to argue that the observed effects are generalizable to the population of participants from which the sample was drawn. Similarly, the researcher selects a few (dozen) stimuli amongst many. S/he is not interested in the selected stimuli *per se* and would like to argue that the observed effects generalize to the population from which the stimuli were drawn. In other words, the researcher would like to be able to conclude that the same effects would be significant if other subsets of participants and stimuli from these same populations were to be selected. For this to be possible, participants *and* stimuli must be treated in statistical models as *random effects*. It is important to note that the statistical terms *random effects* do not imply that the participants or the stimuli were selected at random. This is rarely the case. The mere fact that another researcher might have used other stimuli (as s/he might have used other participants) qualifies stimuli (and participants) as a random effect.

In addition, in most experiments, all participants in a given condition are tested on all stimuli of this condition, and conversely, all stimuli are generally seen by all participants in a given condition. This aspect of the design must also be reflected in the analysis; this is done by treating participant and stimulus as *crossed* random effects. It is important to note that as a consequence of such designs, data sets involve (complex) *repeated measurements*. Each participant is measured several times (on different stimuli) and each stimulus is measured several times (by the different participants). Repeated measurements are undoubtedly characterized by “dependency”. Measurements taken from the same participant are correlated with one another (some participants are faster) and

measurements from the same stimulus are correlated (some stimuli are easier to process than others).

3. Treating stimulus and participant as random effects in the analysis - non ERP data sets

The statistical analysis must treat stimulus and participant as crossed random effects and model the dependency in measurements that arises because the same participants and stimuli are measured multiple times. Random and fixed effects (i.e., those effects accounting for the manipulations/comparisons of interest) must further be modelled jointly (while separating their respective contributions) so as to ensure that the statistical effect of an independent variable can be attributed to the true influence of this variable rather than to specificities of the set of stimuli or participants in the study. This section is concerned with the analysis of experiments in which one data point is collected per trial (e.g., response time, forced choice, response type).

Researchers in cognitive psychology deal/have dealt with these issues in diverse – more or less adequate- ways. The extent to which a given approach is still used varies across fields. In many fields, the dominant approach has long been (or still is) to aggregate the data for each participant in a given condition and to conduct a statistical test (e.g., ANOVA, t-test) on these averaged values to test whether differences across conditions are significant. We will call these analyses classical, by-participant or F1 analyses. Importantly, a significant by-participant analysis allows generalizing the result to the population of participants but only with the same sample of stimuli, and therefore does not provide information on what would happen with another sample of stimuli. The question thus arises of whether the results of these studies can be generalized to other subsets of stimuli and relatedly, of whether some of the discrepancies in findings across studies could be explained by differences in the sets of stimuli used across experiments (see Judd et al., 2012, for examples taken from social psychology, and further discussion).

The necessity to model both participant and stimulus as random effects, especially in language studies, was formalized early on by Clark (1973) following the work of Coleman (1964). Statistical analyses on aggregated psycholinguistic data (other than EEG Data) have since then usually involved

by-participant (F1) and by-stimulus (F2) analyses. In by-stimulus analyses, the data are averaged over participants for each stimulus in a given condition and the analysis is performed on these averaged values. A significant by-stimulus analysis allows generalizing the result to the population of stimuli but only with the same sample of participants, and therefore does not generalize to the population of all possible participants. In order to be able to conclude that the effect would replicate with both a new set of stimuli and a new set of participants, it is necessary to model the two random effects together. Clark offered a solution to this problem which consists in computing a quasi-F statistic (or F'), which is based on the sum of squares derived from the model treating both participant and stimulus as random effects.

Another option to jointly model participant and stimulus as crossed random effects is offered by crossed mixed-effects models. Mixed-effects models (which encompass multi-level or hierarchical models) are an extension of the regression model. Classical regression models are appropriate when each observation is independent from the others. Only one measure per participant can be included. The only random effect is the error term, which can be viewed as a measure of the incapacity of the model to perfectly predict the dependent variable as a function of one or more independent variables. The error is to be understood as a prediction error. If several measures are taken on a same participant (or on the same stimulus) these will be correlated and the model should incorporate this correlation. Mixed-effects models generalize regression models by incorporating variables such as participant or stimulus (random effects) to explain or predict the induced correlation. Random effects account for how the mean of the dependent variable varies with the random variables (i.e., random intercepts) and/or how the influence of the independent variable(s) varies with the random variables (random slopes). The introduction of both random intercepts and random slopes is theoretically required to ensure that the results would generalize to other sets of participants and stimuli (see Barr, Levy, Scheepers, & Tily, 2013), but this is not always possible (or optimal) as it may require more data points than available (e.g., Bates, Kliegl, Vasishth, & Baayen, submitted).

Mixed-effects models are particularly relevant when the researcher wants to introduce continuous predictors in the analysis. Potential confounding variables can be introduced in the model, thereby ensuring that the effects of the predictor of interest can be associated with the researcher's manipulation, i.e., are not (partly) due to other variables. Numerical predictors of interest (i.e., covariates) can further be tested. Unlike the ANCOVA (which also allows introducing continuous predictors) mixed-effects models allow considering, in the same analysis, covariates/confounding variables that describe both, participants (e.g., age, score in an independent task) and stimuli (e.g., complexity of picture, familiarity of face, etc.). In addition, they allow modelling the influence of trial-specific covariates. For instance, predictors accounting for variations in participants' responses over the course of the experiment as a result of fluctuations in attention, fatigue, or strategies, can be brought in the analysis. Doing so often improves the predictive value of the model.

Mixed-models outperform the Quasi-F when data sets are incomplete and the missing data are not completely random. Data sets from experiments in cognitive psychology are characterized by missing data. For instance, in many experiments where the dependent variable is the response time, only correct responses are included in the analysis. By-stimulus, by-participant, and quasi-F analyses may perform poorly when the missing data are not completely at random. They may result in less precision, less power, and biased estimates (Soley-Bori, 2013, see also Lachaud & Renaud, 2011). In many experimental studies, data are not missing completely at random. Often, more data points go missing in the more difficult condition, or for stimuli that are more complex. Mixed-effects models deal well with missing data (Baayen, Davidson, & Bates, 2008) when the variables that account for the missing pattern (e.g., stimulus) can be brought into the statistical model.

Mixed-effects models have spread in some fields of cognitive psychology more than in others. In psycholinguistic research for instance, mixed-effects models with crossed random effects for stimulus and participant have been introduced by Baayen and colleagues about ten years ago (Baayen, 2008; Baayen et al., 2008, see also Lachaud & Renaud, 2011), and their use has since then increased exponentially to become the golden standard for the analysis of responses / response

times (see also Barr et al., 2013; Carson & Beeson, 2013; Janssen, 2012; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). They are spreading to other fields (e.g., Kliegl, Wei, Dambacher, Yan, & Zhou, 2011; Watkins & Martire, 2015) and several researchers advocate a general use of such models in cognitive neuroscience (e.g., Boisgontier & Cheval, 2016; Judd et al., 2012). These models can now be performed in many software (e.g., R: lme4 package, SAS: Proc Mixed; SPSS: Mixed procedure).

3. Treating stimulus and participant as random effects in ERP studies[1]

ERP data sets come with an additional property: they contain multiple measures per trial. Typically, ERPs are locked to the onset of the stimulus and last several hundreds of milliseconds. For each trial, measurements are taken at each time sample, with a sampling rate usually above 250 Hz. For a sampling rate of 512 Hz for instance, 512 measurements are taken per second, which approximately amounts to one measurement every 2 ms. Moreover, for each trial, the EEG signal is often recorded at multiple locations (i.e., electrodes). Studies differ in the number of electrodes on which the signal is measured, from a few electrodes to several hundreds.

The fact that ERPs contain multiple measures per trial poses specific challenges for the statistical analysis. If the researcher wants to analyze more than one measure per trial, s/he will most likely run into the multiple comparison problem. The probability of a Type I error for a single test is typically of 0.05 in psychology experiments (as the significance threshold is set to 5%), but if several tests are conducted and this value is left unchanged across all tests, the family wise error rate (FWER, probability of finding at least one statistical difference under the null hypothesis) increases drastically for each additional test, see for instance von der Malsburg & Angele, 2017, for a recent example on eye-movements in reading studies).

Many ERP experiments in cognitive science select a subset of stimuli amongst many. This aspect of the design is rarely taken into account in the statistical model. Whereas it is well established that

[1] We focus mostly on those frequent cases where the dependent measure of interest is the amplitude of the ERP signal, but the discussion often also applies to other dependent measures.

ERPs show a substantial amount of variability across participants, the inter-stimulus variability is rarely discussed in the literature. To illustrate this variability, we re-analysed the ERP data reported in Bürki (2017). In this study, participants were presented with pictures of objects and were instructed to name these objects aloud. The pictures were presented in four conditions, but the current analysis is restricted to two of them. In the first, the picture appeared with a superimposed written word unrelated to the object's name (neutral condition); in the second, the picture was presented with a line of Xs (no distractor or baseline condition). The data set includes data from 18 participants. We computed a mixed-effects model with random intercepts for participant and stimulus at each time point, with the amplitude of the ERP as the dependent variable, and condition (i.e., no distractor versus neutral distractor) as the fixed effect. Figure 1 displays the random and fixed effects for this analysis over time.

As can be seen on Figure 1, the random intercept for stimulus, although smaller than that for participant, is clearly not negligible. Moreover, it is greater between about 200 and 280 ms after picture onset, a time window which overlaps with the time window in which the estimated parameter associated with the fixed-effect is the greatest (see also Bürki, 2017). Further note that in the present analysis only random intercepts were included. It is possible that participants and stimuli further vary in how their amplitude is influenced by the experimental manipulation. This example illustrates the fact that the amplitude of the ERP signal varies across stimuli. It clearly points to the need to take into account both the variability across participants and stimuli in the analysis.

The optimal analysis of ERP data sets requires that both the fact that stimuli and participants were selected amongst many, and the fact that multiple tests are conducted, be dealt with. How to optimally analyze ERP data sets depends in part on the number of measurements per trial in the time and space dimensions to be considered in the analysis. In the following subsections, we discuss two cases. In the first, the data set contains one (to a few) measures per trial, in the second, the

number of measures is limited in the space dimension but not in the time dimension. Data sets with numerous measures in both dimensions are discussed in section 4.

3.1. One (to a few) measure(s) in the time and space dimensions

ERP data sets often come with hundreds of measurements per trial but it is very common in cognitive psychology to reduce the data set to fewer data points, for instance by averaging the signal in time windows and/or groups of electrodes (e.g., Barkley, Kluender, & Kutas, 2015; Strijkers, Costa, & Thierry, 2010), by performing the analysis on the signal at a given electrode or on the signal averaged across neighboring electrodes (i.e., at a single ROI), by restricting the analysis to specific time points and/or electrode locations (see Keil et al., 2014 on how to (not) select the time window/electrode, or Luck & Gaspelin, 2017)[2] or by focusing on a dependent variable that provides a single measure in the time domain (e.g., maximum of amplitude or its latency[3]). Note that these options[4] come with the need to operate a selection of the time points/time windows (or electrode locations) in which the analysis will be conducted and thus require appropriate a priori knowledge. Without a priori knowledge, the researcher will either select the window(s) arbitrarily or will do so after inspection of the data. In the first case, the windows may be suboptimal and miss an effect that would be maximal at the junction between two time windows. If the selection is driven by a visual inspection of the data, the probability of a Type I error is drastically increased. Spurious effects will draw the attention of the analyst, who will select the time windows and

[2] This might be appropriate when the electrophysiological signature of the phenomenon being examined is well defined (e.g., meaning processing indexed with the N400 component, Kutas & Federmeier, 2011; processing of syntactic abnormalities with the P600, e.g., Gouvea, Phillips, Kazanina, & Poeppel, 2010; face processing/familiarity indexed with the N170, e.g., Alonso-Prieto, Pancaroglu, Dalrymple, Handy, Barton, & Oruc, 2015). In many cases, however, it is difficult to optimally select the time points and/or electrodes a priori, because participants tend to differ in the timing of components (see for instance Alonso-Prieto et al., 2015) and the averaging of the signal over time, despite its disadvantages, might be better suited than the selection of time points/electrodes for this reason.

[3] This approach is particularly suited when the researcher targets a specific component, as the peak of amplitude will reflect the peak of this component. It is also particularly appropriate to deal with inter-individual differences in the timing of components (e.g., Gaspar et al., 2011).

[4] A fourth way of reducing the data set is to use Independent Component Analysis (ICA). The purpose of ICA is to separate linearly mixed sources. This technique is often used to separate the signal of interest from artefacts but can also be used to decompose the signal in a sum of components (https://sccn.ucsd.edu/~scott/PNAS.html) and the statistical analysis can then be performed on these components (e.g., Groppe, Makeig, & Kutas, 2009; Onton & Makeig, 2006).

locations in order to maximize the chances of finding a significant difference. Selecting the best window visually is equivalent to testing the effect on all windows and reporting only the most significant one as if it was the only one that was considered a priori. In short, the need to select time windows for the analysis decreases the chances that the results will be replicated or can be compared. Several recent studies have shown the poor replicability of psychological experiments. According to Simmons and colleagues (2011) this lack of replicability can be explained by the many degrees of freedom that the researcher has during the analysis of the data. The researcher's degrees of freedom increase drastically if time windows and locations are selected a posteriori. Analyses that limit the number of arbitrary decisions (e.g., mass univariate analyses, see below) are a much better option in this respect.

Analyses on reduced data sets were required when computational resources did not allow processing big databases. With contemporary computational resources, however, this is no longer the case. In practice, this is still done often, and may be justified by the research question, or a priori knowledge.

Statistical analysis of reduced data sets

When the data set is reduced to contain, say only one measure in the time dimension (with or without electrode location as an independent variable), the statistical models discussed previously, which are appropriate for the analysis of data sets with reaction times or error rates, are also optimal in this case. Surprisingly, however, the statistical models that are used in most ERP studies dealing with such data sets are inadequate in that they do not treat stimulus as a random variable. In most cases, a by-participant t-test or ANOVA is performed (i.e., F1 analysis) only. Once again, with this approach, the researcher can conclude that the results would replicate with a different set of participants using the same set of stimuli, but cannot conclude that they generalize to different sets of stimuli (see Clark, 1973; or Judd et al., 2012).

Forster and Dickinson (1976) showed that Type I error rates for analyses that do not consider stimulus as a random effect can exceed the alpha level by a factor of 10 (see also Lachaud & Renaud,

2011). Here, we present additional evidence. We ran Monte-Carlo simulations to compute the percentages of cases in which the null hypothesis is rejected for simulated[5] data sets of 20 participants and 18 or 36 stimuli, with three factors; A (between participants, within stimuli), B (within participants, between stimuli), and C (within participants, within stimuli), and their interactions[6] under the null hypothesis (see Table I). We compared the type I error rate for a nominal level of 0.05 for the Quasi-F statistic, the classical ANOVA (F1), and the crossed mixed-effects model (see Lachaud & Renaud, 2011 for a similar approach with comparisons between different F statistics, amounts of missing data, replacement schemes, and distributions). Simulations are run on the R language, based on the permuco package (Frossard & Renaud, 2018) for the F1 and Quasi-F statistics, and on the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) for the mixed-effects model statistic (Kenward-Roger statistic). The results are summarized in Table I. The code for all functions involving the F1 statistic in this paper can be found in the permuco package (available here https://cran.r-project.org/package=permuco) and functions involving the Quasi-F statistic can be found on a github repository (https://github.com/jaromilfrossard/permucoQuasiF, that will eventually be merged into permuco).

Appropriate methods should result in values that are close to 0.05. The obtained proportions for the Quasi-F and mixed-effect models are quite close to this value, which indicates that their *p*-values are trustworthy. By contrast, the classical F1 approach has an unacceptable type I error rate, that rockets up to 60%. This means that in this setting, 60% of no-effects are declared significant. Anticipating the need of permutation for the next section, Table I also provides the results of two

[5] All our simulations were conducted on artificial (simulated) datasets

[6] 4000 samples were generated under the null hypothesis. The data are simulated using random effects with all possible intercepts and slopes and with a normal distribution, and with decreasing variability with respect to the interaction level (for random intercepts $\sigma=1$, for random slopes $\sigma=1/2$, for random interaction of level 2 $\sigma=1/3$). The variability of the error terms is $\sigma=2$.

permutation approaches, in which the *p*-value is obtained through permutation[7]. As expected given that the requirements of the parametric approach are fulfilled, the proportion of rejected null hypotheses is quite similar for the parametric and permutation approaches.

This simulation adds to the many demonstrations in the literature that the use of methods that do not take into account the effects of the stimuli have type I error rates that are much larger than expected. It can thus be hypothesized that a non-negligible number of the significant results reported in the literature are in fact driven by specificities in the set of stimuli selected for the experiment and would not replicate with different sets of stimuli, fostering the replicability crisis.

Nothing prevents researchers to perform by-stimulus analyses, quasi-F analyses, or mixed-effects models with reduced ERP data sets (see for instance Bedny, Aguirre, & Thompson-Schill, 2007 for by-stimulus analyses of fMRI data). Mixed-effects models with the signal in the selected time window, a selected time point, or the peak of amplitude as the dependent variable, the predictors of interest and random effects for participant and stimulus have been used in a few papers (Khalifian, Stites, & Laszlo, 2016; Payne, Lee, & Federmeier, 2015, see also, Amsel, 2011; Toscano, McMurray, Dennhardt, & Luck, 2010; Vossen, Van Breukelen, Hermens, Van Os, & Lousberg, 2011 for other types of mixed-effects models) but their use is far from frequent. The software R (R development core team, 2015) and library lme4 (Bates et al., 2015, see also lmerTest, Kuznetsova,

[7] For the classical ANOVA, we used the method from Kherad-Pajouh & Renaud (2015). For the quasi-F, we used the method from ter Braak (1992), where all the estimated random effects are randomly flipped to obtain new data. More precisely, for the Quasi-F, random and fixed effects were estimated as if they were all fixed effects. For the random effects, we did not estimate the parameters (which are constrained to sum to zero), but a choice of orthogonal contrasts (which are free). We kept the fixed part and flipped the contrasts of the random part randomly to form new permuted dependent variables. The F statistics were transformed to log p-values, to anticipate the generalisation of the method to ERP data (see section 3.2) and more specifically to allow the use of Fisher's P-value combining method, which sums the logarithms of the p-values (Smith & Nichols, 2009).

Brockhoff, & Christensen, 2017, and afex, Singmann, Bolker, Westfall, & Aust, 2017) can for instance be used to run such models. Once again, mixed-effects models offer more flexibility than quasi-F statistics and are to be preferred when the data set is unbalanced, or when it makes sense to include continuous predictors in the analysis. By now, they should have become the golden standard with ERP reduced data sets.

Multiple testing issue

A frequent practice consists in averaging the signal in several successive time windows (or intervals), and to run a statistical analysis *separately* in each time window. Conclusions on the time course of the effect are based on the presence/absence of the effect in these time windows. Hence multiple tests are performed and the multiple comparison problem must be addressed. Yet, most studies do not correct for multiple comparisons in this context.

The most popular and best known method to correct for multiple comparisons is the Bonferroni correction. It consists in dividing the nominal level alpha by the number of comparisons, thereby ensuring that the FWER does not exceed the alpha level. The Bonferroni correction is however typically too conservative when the tests are not independent, as is the case with ERPs. Using the Bonferroni correction in this context would result in a decrease in statistical power (probability of correctly rejecting the null hypothesis). Different options that take into account the correlation between the tests while maintaining the FWER have been proposed, either using the closure principle or a resampling approach (see for instance Bretz, Hothorn, & Westfall, 2016 for details) and are available in R in the multcomp (Hothorn, Bretz, & Westfall, 2008) and coin packages (Hothorn, Hornik, van de Wiel, & Zeileis, 2008). Their description goes beyond the scope of this article, but note that these procedures offer substantial improvements compared to the Bonferroni procedure in terms of power. Another criterion to correct for multiple comparisons is the false discovery rate (FDR, or the expected proportion of falsely rejected hypotheses among the rejected hypotheses). This criterion has militants and objectors (see for instance Bretz et al., 2016). One of its advantages is that there exists a relatively powerful and easy method to control the FDR, called

Benjamini-Hochberg (Benjamini & Hochberg, 1995). Note, however, that this method does not control the FDR in all cases (of correlation structure). Again, we refer to Bretz et al. (2016) for more information.

3.2. One (or a few) electrodes, many time points

Whether it is appropriate (or optimal) to reduce the data set in the temporal dimension depends on the study, the amount of background knowledge, and the purpose of the analysis. Whereas data reduction can be appropriate when the researcher has good a priori knowledge about the timing of an effect, it is not in all cases where this is not the case. The selection of time windows is for instance not appropriate when the purpose of the study is precisely to determine the time course of an effect. Testing for an effect in successive time windows in which the signal has been averaged only provides a very rough temporal measure of the boundaries of an effect. If the time windows of analyses are for instance 100-249, 250-399 and 400-500 ms after stimulus onset, and an effect is found in the latter only, it is not possible to know whether the onset of the effect is 400, or, say, 480.

In many cases, the appeal of ERPs is specifically that they have a high temporal resolution. A high number of measurements, in the time domain at least, is highly desirable for these studies and the data are most informative if analyzed with as little a priori restrictions in this dimension. An increasingly popular statistical approach is to perform the statistical modeling on the unaveraged data (also sometimes called single trial analysis). One statistical test is performed at each time point and specific techniques are used to correct for multiple comparisons.

Addressing the multiple comparison problem appropriately requires accounting for the dependency in measurements in the time dimension. Some have dealt with this by deciding on the number of consecutive time frames that will have to show the effect, at a given alpha level, for the effect to be considered significant. This approach has to be preferred over a selection based on data inspection or any other double-dipping strategy (e.g., selecting a time window for an analysis based on an analysis which is not independent of the first, see Kriegeskorte, Simmons, Bellgowan, & Baker,

2009) but is not optimal. It requires the interval size to be defined arbitrarily. To illustrate this problem, imagine you decide for 20 ms and the effect lasts 19 seconds. What should be done? Moreover the error rate of this procedure -- its FWER -- is not known and depends on the characteristics of the data (essentially its correlation structure).

Addressing the multiple comparison problem

As mentioned above, different methods have been proposed to correct for the multiplicity of tests. This is also true in the context of non-aggregated ERP data sets. The comparison of these approaches is beyond the scope of the present paper. An excellent review with pros and cons can be found in Groppe, Urbach, and Kutas (2011). The different methods vary in the criterion they aim at (e.g., FWER, FDR) and the path they take to achieve this criterion. As underlined by Groppe et al. (2011), a key aspect in which these methods differ is "how they negotiate the trade-off between statistical power (…) and the degree of certainty one can have in the significance of any single test result" (Groppe et al., 2011, p.18.). One such method, cluster mass tests, has recently become popular in the EEG community.

Cluster mass tests offer an elegant solution to the multiple comparison problem and formally control weaky the FWER. They rely on cluster mass statistics (see for instance Maris & Oostenveld, 2007 and Cheval et al., 2018, or Luque et al., 2017, for selected examples in the analysis of ERPs). A statistics (e.g., a *t* or a *F statistics*) is computed at each time point, but instead of computing a *p*-value for each of these time points, all the time points that exhibit a statistics higher than a given threshold are grouped together, clustered based on their temporal (or spatio-temporal, see below) adjacency. Cluster mass tests require that a threshold be defined a priori. In practice, the threshold is often the 95% quantile of the F distribution. For each cluster, a cluster statistic, the cluster mass, is computed based on the statistics of each of the time points it contains. A permutation scheme evaluates the significance of each cluster, as a whole. Different cluster-level statistics can be used depending on the researcher's a priori knowledge about the effect, i.e., largely distributed in time and/or space or rather local (see for instance Pernet, Chauveau, Gaspar, & Rousselet, 2011). An

extension of the cluster-mass test that avoids the choice of a threshold while allowing different *p*-values for each time point has been proposed by Smith & Nichols (2009), the Threshold Free Cluster Enhancement, or TCFE. In this approach, the height of the cluster (maximal value) and its extent (number of elements) are integrated for all possible thresholds. The way the cluster extent and height are used in the integral (power function) must be decided a priori (see Pernet, Latinus, Nichols & Rousselet, 2015 for consensus values).

This method, like other ways of correcting for multiple comparisons, has limitations (we again refer the reader to Groppe et al., 2011 for an extensive discussion) but also many advantages. It is for instance very powerful to detect effects that are broadly distributed in time (permutation tests based on single *t* values -rather than on cluster-based statistics- might for instance be more appropriate for detecting narrow effects or obtain more precise information on the onset and offset of an effect). Notably, cluster mass tests in many articles (including Maris & Oostenveld, 2007) and many software implementations allow for only one factor (typically 2 conditions). This is due to the difficulty of defining a correct permutation scheme in cases with more factors, as illustrated next. To test the null hypothesis of no effect between the two modalities of a factor, say A, one can freely permute the data points since the data share the same distribution under the null hypothesis. However, in the presence of a second factor, say B, the null hypothesis about A does not impose the data from the two levels of B to have the same mean. This implies that to test the null hypothesis of A, one cannot freely permute the data anymore. This point is often overlooked in many uses of permutation tests. Several generalizations of the basic permutation scheme that account for this difficulty have been proposed (see Kherad-Pajouh & Renaud, 2010; Kherad-Pajouh & Renaud, 2015; or Winkler, Ridgway, Webster, Smith, & Nichols, 2014). Note that in the simplest case, the permutation test is exact, meaning that it rejects exactly 5% of data under the null hypothesis when the alpha level is set to 5%. None of the extensions discussed above keeps this property for finite samples but they are usually close (Anderson, 2001; Winkler et al., 2014). Related to permutations, the bootstrap could also be considered. It is never exact, but when both

permutation and bootstrap are adequate, they tend to give similar results, as shown for example in Pernet et al. (2015) for ERPs.

As discussed in section 2, in experiments with participants and stimuli selected amongst many, we want the statistical test to treat both stimulus and participant as random effects, and to do so simultaneously. We stressed that the Quasi-F and statistics from the mixed-effects models are the best options, and highlighted the additional advantages of the mixed-effects model approach. To our knowledge, cluster mass statistics have not yet been used with the Quasi-F or statistics from mixed-effects models. The only requirement is to obtain statistics at each time point that are aggregable and a permutation method that generates ERPs that are coherent with the null hypothesis. The permutation-based Quasi-F method proposed in section 3.1 can therefore be used with the cluster mass procedure. In the next section, we present simulations combining the Quasi-F and cluster mass tests. For comparison purposes, we also combine the Quasi-F statistic with the Benjamini-Hochberg (BH) procedure. Note that the package permuco (Frossard & Renaud, 2018) and its extension for the Quasi-F statistic (https://github.com/jaromilfrossard/permucoQuasiF, that will eventually be merged into permuco) provides functions to apply these two methods as well as other methods, which are not detailed in the present study.

Comparing FWER obtained with cluster mass tests for F1 and Quasi-F statistics

In this section we present simulations that compare the FWER obtained with cluster mass tests and the BH method when used with the (classical) F1 statistic versus with the Quasi-F. The most important results are shown in Table II.

For the simulations of ERPs, as in section 3.1, we simulated data sets of 20 participants with 18 or 36 stimuli, with three factors: A (between participants, within stimuli), B (within participants, between stimuli), and C (within participants, within stimuli), and their interactions. We simulated

an ERP of 600 time points for each participant exposed to each stimulus[8]. For cluster-mass based tests, at each iteration, a cluster-mass test was performed for the main effects of A, B, and C using either the by-participant F1 statistic or the quasi-F statistic. The table summarizes the average FWER for these two methods (i.e., proportion of samples that have at least one significant cluster). By definition, a correct method should result in proportions close to the nominal value of 0.05. The quasi-F method is not perfect as the proportion oscillates between 6% and 9%. The rate is much worse however for the F1 statistic (between 8% and 100%). For instance, a significant cluster is found for the main effect of B in 100% of the iterations with 18 stimuli and in 99.7% of the iterations with 36 stimuli (instead of the expected 5%)! These results are even more alarming than in section 3.1 and truly question the replicability of ERP studies analyzed with F1 statistics. Note that with a fully balanced design, without covariates, and the same permutation scheme, the results of the simulations would most probably be the same with a statistic (e.g., Kenward-Roger) from a mixed-effects model with the maximal random effect structure (see Barr et al., 2013), but the computation time would be much greater than with the Quasi-F.

Simulations with the BH procedure show that as expected, the procedure performs poorly when applied to the F1 *p*-value (Classical F1, BH in Table II) as it does not take into account the variance induced by the stimuli. By contrast, when this variance is brought into the model, as in the Quasi-F (Quasi-F, BH) the results are quite good, although a little bit conservative.

[8] 4000 samples of ERPs were artificially generated using the sum of signals due to the participant random effect, the stimulus random effect, their interactions with fixed effects (random slopes) and the errors. To obtain smooth curves similar to ERPs, each of these signals was generated as a multivariate Gaussian vector using a covariance matrix that decays very slowly as a function of time distance, while still being positive definite. More precisely, the temporal correlations were simulated using an exponential correlation function $\rho(\tau) = \exp(-3\tau^2 / R)$ (Abrahamsen, 1997), also called Gaussian correlation function. The range parameter was set to R=60 for the random effects associated with the participants, to R=40 for the random effects associated with the stimuli and to R=20 for the error terms. As in section 3.1, the standard deviations (amplitude) are decreasing with the order of the interactions.

The results of additional simulations on different sample sizes (for participants and for stimuli) can be found in the supplementary material. These simulations all show that the stimuli have to be treated as a random variable in the statistical model. An interesting goal for future work will be to compare the results of different methods or combination of methods that treat stimulus as a random variable and correct for the multiplicity of tests as Groppe et al. (2011b) did for by-subject statistics. Whereas the choice of the statistic (e.g., Quasi-F versus statistic form the mixed-effect model[9]) and the selected method to correct for the multiplicity of tests can be expected to show small variations in the FWER, it is clear from the present simulations that methods that do not account for the variance induced by the stimuli are to be avoided.

Power considerations

Power analyses are presented in the supplementary material, for the two datasets included in the simulation described above, as well as for other simulated datasets with different numbers of stimuli and participants. Since the classical F1 approach does not control the FWER, the power between the F1 and Quasi-F approaches should not be compared, and researchers should not use methods that are known to be anti-conservative in any case. It is however reassuring to see that the power of the F1 and Quasi-F are relatively similar. This suggests that using the Quasi-F approach does not imply a high cost in statistical power.

4. Current challenges in treating stimulus as a random effect in the analysis of ERPs

We discussed cases where it is reasonable to reduce the data set to one (or a few more) measures. The approach to be adopted here is straightforward and already used with dependent variables such as reaction times, or response type. Our simulation adds to other contributions in the literature showing that failure to adopt this approach leads to non-negligible increases in type I

[9] The use of mixed-effects model in this context would be particularly relevant, but several points will have to be resolved such as which random effects to include in the random effect structure of the model, or how to cope with a lack of convergence of the models at some time points.

error rates. We then discussed cases in which it is reasonable to have only one (or a few electrodes) and the researcher wants the analysis to tell when an effect starts and ends, or simply whether there is an effect, without having to choose (a) specific time window(s). We showed that it is possible to conjugate cluster mass tests with Quasi-F and demonstrated, using simulation, the advantage of this approach over a by-participant analysis. A first challenge for future research will be to determine whether this approach can be used with mixed-effects models when the data sets are unbalanced or when covariates should be taken into account. This requires that a permutation scheme be defined that respects the structure of the data and is likely to be quite computationally intensive.

A second challenge for the statistical analysis of ERPs consists in treating stimulus as a random variable in data sets that are unreduced in both the time and space dimensions. In some cases, it may indeed be desirable to record multiple measures in both the time and space dimensions. This is especially true when the researcher does not know where and when the effect is to be expected. The cognitive processes of interest are not always indexed by specific components and the purpose of the ERP analysis might specifically be to describe the time course and location of the effect of an experimental manipulation. It may not be a good idea in this context to select electrodes a priori or average the signal on a subset of electrodes (more generally, see Brooks, Zoumpoulaki, & Bowman, 2017 on the dangers of selecting regions of interest; or Dien, 2017 on averaging across electrodes). *Mass univariate analyses* serve exactly this purpose. They consist in performing one statistical test for each time point and location (e.g., Groppe et al., 2011; Pernet, et al., 2011). Pernet and collaborators (2011) implemented for instance cluster mass permutation tests in the context of a two-level mass univariate analysis (an approach commonly used to analyze fMRI data). At the first level, the analysis is performed for each participant individually, and consists in a General linear model performed independently at each time point and electrode. The output of each model is a Beta coefficient for each condition. At the second level of analysis, the coefficients obtained at the first level are analyzed across participants, to assess significance at the group level. The exact model

performed at this stage depends on the data and research question. Significance is assessed using the cluster mass method. Interestingly, the same approach can be used to assess significance at the participant level (see for instance Salvia et al., 2014 for an example in face recognition research). The approach described by Pernet and collaborators has been used in a few domains so far, including the study of early visual evoked potentials; face recognition Alonso-Prieto et al., 2015; letter recognition, Madec et al., 2016; or language production research, Bürki, Sadat, Dubarry, & Alario, 2016).

The two-step procedure in Pernet et al. can be seen as a hierarchical multilevel model, a special case of mixed-effects model. The first-level analysis is performed at the single trial level, therefore modeling the variability at this level. The second level models the variability across participants. In this approach however and unlike in mixed-effects models with crossed random effects for participant and stimulus, stimulus is not modeled as a random effect. Rather, trials are modeled as a random effect nested under participants. In addition, trial and participant are modeled separately rather than jointly. This analysis thus does not model the dependency that arises because the same stimuli are seen by different participants and as a consequence, missing data may not be dealt with in the most optimal way. Note that this two-step procedure has also been described (and used) by others under different names, as reviewed for instance in Smith and Kutas (2015). In most cases, however, the two-level model was not used in conjunction with the cluster mass approach to address the multiple comparisons problem.

An obvious question is whether mass univariate analyses with cluster mass tests can be combined with a method that treat stimulus as a random effect. To our knowledge, the use of Quasi-F analyses or crossed mixed-effects models with genuine cluster-mass tests has not been implemented for spatio-temporal clusters in regular software. Although heavier computationally, the permutation methods presented in section 3 could be used with unreduced data sets. The availability of

increasing computational resources will likely make it possible in a near future to adapt the cluster mass test for Quasi-F in the context of mass univariate analyses.[10]

Another option to analyze data sets which are unreduced in the time dimension is found in Generalized additive mixed-effects models (GAMMs, see Hendrix, Bolger, & Baayen, 2017; Kryuchkova, Tucker, Wurm, & Baayen, 2012; Meulman, Wieling, Sprenger, Stowe, & Schmid, 2015). A GAMM is a non-parametric generalized linear model in which the linear predictors depend, in part, on a sum of smooth functions of predictors. Such models can for instance be performed in the mgcv R package (Wood, 2018, see also the itsadug package for visualisation, Van Rij, Wieling, Baayen, & van Rijn, 2015). For instance, Hendrix et al. (2017) fitted independent GAMMs for each of the 32 electrodes at which the signal was recorded. They included by-participant factor smooths for trial and time and random intercepts for their stimuli in the random effect structure of the model. In the fixed-part, they included a main effect smooth for each predictor of interest, and a tensor product for the interactions between each predictor and time. Given that time is included in the analysis, there is a priori no need to perform multiple analyses in the time dimension. The researcher is however often interested in determining the time window in which the effects are significant. To obtain such information, one could compute confidence intervals at each time point, and look at the first point in time when zero is not in the interval. With this procedure, the multiple comparison problem is back again, given that multiple tests (i.e., the confidence intervals) are conducted independently at each time point. An alternative option would be to compute joint

[10] A few studies used mass univariate mixed-effects models (one model at each time point and electrode) but without cluster mass tests to correct for multiple comparisons. For instance Janssen, Hernández-Cabrera, van der Meij, & Barber (2015) performed multiple mixed-effects models at each time point and electrode. They considered an effect to be significant if the *p* value was below 0.05 during at least 48 ms. The result of the first analysis was used to define the time window in which a second analysis was to be performed (see here Kriegeskorte et al., 2009, on the consequences of what they call "double-dipping" strategies). Note also that the random effect structure of the models in this study only included random intercepts.

intervals. In such cases, significance can be assessed for each individual data point, using the joint intervals. It must be noted however that joint intervals are much less powerful that the cluster mass to assess the significance of individual data points.

GAMMs allow relaxing the assumption that the predictors are linearly related to the dependent variable. Results of GAMMs may however not be as easy to interpret as results from linear mixed-effects models and some authors view GAMMs as more suited as an exploratory tool than as an inference tool (e.g., Fink & Hochachka, 2009; Xiang, 2001).

GAMMs have so far dealt with data sets that are unreduced in the time dimension, but not in the space dimension. Studies either selected one (or a subset of) electrode(s) or averaged the signal over several electrodes. As highlighted by Meulman et al. (2015), it is theoretically possible to include many electrodes in a GAMM analysis but this would require higher computational resources than currently available.

5. Conclusion

ERP data from psychological experiments are complex data. They are taken from selected subsets of participants and stimuli, with several measures taken per participant and stimulus. In addition, thousands of measures are taken at each trial. An accurate analysis of such data requires that this complexity be taken into account. The majority of ERP studies in cognitive science use inadequate analyses, questioning the reliability of published findings. The number of studies using ERPs has increased drastically in the last ten years and this trend will likely continue. There is an urgent need to think about statistical practices and how they can be improved. Failure to do so inflates family-wise error rates, feeding the replication crisis in cognitive science/psychology.

Crucially, most current analyses do not allow a generalization of the findings to other sets of stimuli. The need to treat stimulus as a random effect in psychology was highlighted many decades ago (e.g., Clark, 1973). The same applies to ERP studies but despite this fact, most analyses do not deal with this issue at all. In this contribution, we discussed several options. When the analysis is

performed on a reduced data set (selection of / averaging across specific time intervals/locations) researchers can use Quasi-F analyses or crossed mixed-effects models (especially with missing data) instead of the classical by-participant analysis. This way, they can conclude that the effect would replicate with a different set of participants processing a different set of stimuli. For data sets reduced in the space but not in the time dimension, which are by far the most frequent in cognitive psychology, recent developments in statistical software and packages make it possible to use the Quasi-F statistic. This approach now needs to be extended to data sets that are unreduced in both the time and space dimensions.

Supplementary material

Refer to web version for supplementary material

Declarations of interest

None

Acknowledgments

This work was partly supported by the Deutsche Forschungsgemeinschaft (DFG), Collaborative Research Centre SFB 1287, Project B05. The computations were performed at the University of Geneva on the Baobab cluster.

References

Abrahamsen, P. (1997). A review of Gaussian random fields and correlation functions. Retrieved from https://www.nr.no/directdownload/917_Rapport.pdf

Alonso-Prieto, E., Pancaroglu, R., Dalrymple, K. A., Handy, T., Barton, J. J. S., & Oruc, I. (2015). Temporal dynamics of the face familiarity effect: Bootstrap analysis of single-subject event-related potential data. Cognitive Neuropsychology, 32, 266–282. https://doi.org/10.1080/02643294.2015.1053852

Amsel, B. D. (2011). Tracking real-time neural activation of conceptual knowledge using single-trial Event-Related potentials. Neuropsychologia, 49, 970–983. https://doi.org/10.1016/j.neuropsychologia.2011.01.003

Anderson, M. J. (2001). Permutation tests for univariate or multivariate analysis of variance and regression. Canadian journal of fisheries and aquatic sciences 58, 3, 626-639.

Baayen, R. H. (2008). Analysing linguistic data. A practical introduction to statistics using R. New York: Cambridge University Press.

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. https://doi.org/10.1016/j.jml.2007.12.005

Barkley, C., Kluender, R., & Kutas, M. (2015). Referential processing in the human brain: An Event-Related Potential (ERP) study. Brain Research, 1629, 143–159. https://doi.org/10.1016/j.brainres.2015.09.017

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., Kliegl, R., Vasishth, S., & Baayen, H. (Submitted). Parsimonious Mixed Models. arXiv:1506.04967 [Stat]. Retrieved from http://arxiv.org/abs/1506.04967

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67, 1 - 48. doi:http://dx.doi.org/10.18637/jss.v067.i01

Bedny, M., Aguirre, G. K., & Thompson-Schill, S. L. (2007). Item analysis in functional magnetic resonance imaging. NeuroImage, 35, 1093–1102. https://doi.org/10.1016/j.neuroimage.2007.01.039

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. Retrieved from http://www.jstor.org/stable/2346101

Boisgontier, M. P., & Cheval, B. (2016). The anova to mixed model transition. Neuroscience and Biobehavioral Reviews, 68, 1004–1005. https://doi.org/10.1016/j.neubiorev.2016.05.034

Bretz, F., Hothorn, T., & Westfall, P. (2016). Multiple Comparisons Using R. CRC Press.

Brooks, J., Zoumpoulaki, A., & Bowman, H. (2017). Data-driven region-of-interest selection without inflating Type I error rate: Safe data-driven ROI selection. Psychophysiology, 54, 100–113. https://doi.org/10.1111/psyp.12682

Bürki, A. (2017). Electrophysiological characterization of facilitation and interference in the picture-word interference paradigm. Psychophysiology, 54, 1370-1392. . https://doi.org/10.1111/psyp.12885

Bürki, A., Sadat, J., Dubarry, A.-S., & Alario, F.-X. (2016). Sequential processing during noun phrase production. Cognition, 146, 90–99. https://doi.org/10.1016/j.cognition.2015.09.002

Carson, R. J. & Beeson, C. L. M (2013). Crossing language barriers: Using crossed random effects modelling in psycholinguistics research. Tutorials in quantitative methods for psychology, 9, 25-41. https://doi.org/ 10.20982/tqmp.09.1.p025

Cheval, B., Tipura, E., Burra, N., Chanal, J., Orsholits, D., Radel, R., Boisgontier, M. P. (2018). Avoiding sedentary behaviors requires more cortical resources than avoiding physical activity: An EEG study. BioRxiv 2018. doi: 10.1101/277988

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

Coleman, E. B. (1964). Generalizing to a language population. Psychological Reports, 14, 219-226.

Dien, J. (2017). Best practices for repeated measures ANOVAs of ERP data: Reference, regional channels, and robust ANOVAs. International Journal of Psychophysiology, 111(Supplement C), 42–56. https://doi.org/10.1016/j.ijpsycho.2016.09.006

Fink, D. & Hochachka, W. (2009). Gaussian semiparametric analysis using hierarchical predictive models. In D. L. Thomas, E. G. Cooch, and M. J. Conroy (Eds.), Modeling demographic processes in marked populations. New-York: Springer.

Forster, K., & Dickinson, R. G. (1976). More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F1, F2, F', and min F'. Journal of Verbal Learning and Verbal Behavior, 15, 135–142. https://doi.org/10.1016/0022-5371(76)90014-1

Frossard J., & Renaud O. (2018). permuco: Permutation Tests for Regression, (Repeated Measures) ANOVA/ANCOVA and Comparison of Signals. R package version 1.0.0. https://CRAN.R-project.org/package=permuco

Gaspar, P. A., Ruiz, S., Zamorano, F., Altayó, M., Pérez, C., Bosman, C. A., & Aboitiz, F. (2011). P300 amplitude is insensitive to working memory load in schizophrenia. BMC Psychiatry, 11, 29. https://doi.org/10.1186/1471-244X-11-29

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. Language and Cognitive Processes, 25, 149–188. https://doi.org/10.1080/01690960902965951

Groppe, D. M., Makeig, S., & Kutas, M. (2009). Identifying reliable independent components via split-half comparisons. Neuroimage, 45(4), 1199–1211. https://doi.org/10.1016/j.neuroimage.2008.12.038

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. Psychophysiology, 48(12), 1711–1725. https://doi.org/10.1111/j.1469-8986.2011.01273.x

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011b). Mass univariate analysis of event-related brain potentials/fields II: Simulation studies. Psychophysiology, 48(12), 1726–1737. https://doi.org/10.1111/j.1469-8986.2011.01272.x

Hendrix, P., Bolger, P., & Baayen, H. (2017). Distinct ERP signatures of word frequency, phrase frequency, and prototypicality in speech production. Journal of Experimental Psychology: Learning, Memory, and Cognition, 43, 128-149. https://doi.org/10.1037/a0040332

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. Biometrical Journal 50(3), 346--363. https://doi.org/10.1002/bimj.200810425

Hothorn, T., Hornik, K., van de Wiel, M. A., & Zeileis, A. (2008). Implementing a Class of Permutation Tests: The coin Package. Journal of Statistical Software 28(8), 1-23. URL http://www.jstatsoft.org/v28/i08/.

Janssen, D. P. (2012). Twice random, once mixed: Applying mixed models to simultaneously analyze random effects of language and participants. Behavior Research Methods, 44, 232–247. https://doi.org/10.3758/s13428-011-0145-1

Janssen, N., Hernández-Cabrera, J. A., Meij, M. van der, & Barber, H. A. (2014). Tracking the time course of competition during word production: Evidence for a post-retrieval mechanism of conflict resolution. Cerebral Cortex, bhu092. https://doi.org/10.1093/cercor/bhu092

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, 54–69. https://doi.org/10.1037/a0028347

Keil, A., Debener, S., Gratton, G., Junghöfer, M., Kappenman, E. S., Luck, S. J., … Yee, C. M. (2014). Committee report: Publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. Psychophysiology, 51, 1–21. https://doi.org/10.1111/psyp.12147

Khalifian, N., Stites, M. C., & Laszlo, S. (2016). Relationships between event-related potentials and behavioral and scholastic measures of reading ability: A large-scale, cross-sectional study. Developmental Science, 19, 723–740. https://doi.org/10.1111/desc.12329

Kherad-Pajouh, S., & Renaud, O. (2010). An exact permutation method for testing any effect in balanced and unbalanced fixed effect ANOVA. Computational Statistics & Data Analysis, 54, 1881–1893. https://doi.org/10.1016/j.csda.2010.02.015

Kherad-Pajouh, S., & Renaud, O. (2015). A general permutation approach for analyzing repeated measures ANOVA and mixed-model designs. Statistical Papers, 4, 947–967.

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. Frontiers in Psychology, 1. https://doi.org/10.3389/fpsyg.2010.00238

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. Nature Neuroscience, 12, 535–540. https://doi.org/10.1038/nn.2303

Kryuchkova, T., Tucker, B., H Wurm, L., & Baayen, H. (2012). Danger and usefulness are detected early in auditory lexical processing: Evidence from electroencephalography. Brain and language, 122, 81–91. https://doi.org/10.1016/j.bandl.2012.05.005

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). Annual Review of Psychology, 62, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. Journal of Statistical Software, 82, 1 - 26. http://dx.doi.org/10.18637/jss.v082.i13

Lachaud, C. M., & Renaud, O. (2011). A tutorial for analyzing human reaction times: How to filter data, manage missing values, and choose a statistical model. Applied Psycholinguistics, 32, 389–416. https://doi.org/10.1017/S0142716410000457

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). Psychophysiology, 54, 146–157. https://doi.org/10.1111/psyp.12639

Luque. D., Beesley, T., Morris, R. W., Jack, B. N., Griffiths, O., Whitford, T. J., & Le Pelley, M. E. (2017). Goal-directed and habit-like modulations of stimulus processing during reinforcement learning. Journal of Neuroscience, 37 (11), 3009-3017. doi: 10.1523/JNEUROSCI.3205-16.2017

Madec, S., Goff, K. L., Riès, S. K., Legou, T., Rousselet, G., Courrieu, P., … Rey, A. (2016). The time course of visual influences in letter recognition. Cognitive, Affective, & Behavioral Neuroscience, 16, 406–414. https://doi.org/10.3758/s13415-015-0400-5

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods, 164, 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

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

Meulman, N., Wieling, M., Sprenger, S. A., Stowe, L. A., & Schmid, M. S. (2015). Age effects in L2 grammar processing as revealed by ERPs and how (not) to study them. PLOS ONE, 10(12), e0143328. https://doi.org/10.1371/journal.pone.0143328

Onton J., & S. Makeig. (2006). ICA example and Overview chapter: Why use ICA to decompose EEG/MEG data? Progress in Brain Research. https://doi.org/10.1016/S0079-6123(06)59007-7

Payne, B. R., Lee, C.-L., & Federmeier, K. D. (2015). Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. Psychophysiology, 52, 1456–1469. https://doi.org/10.1111/psyp.12515

Pernet, C. R., Chauveau, N., Gaspar, C., & Rousselet, G. A. (2011). LIMO EEG: A Toolbox for Hierarchical LInear MOdeling of ElectroEncephaloGraphic Data. Computational Intelligence and Neuroscience, 2011, e831409. https://doi.org/10.1155/2011/831409

Pernet, C. R., Latinus, M., Nichols, T. E., & Rousselet, G. A. (2015). Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: A simulation study. *Journal of Neuroscience Methods*, *250*(Supplement C), 85–93. https://doi.org/10.1016/j.jneumeth.2014.08.003

van Rij, J., Wieling, M., Baayen, R. & van Rijn, H. (2017). "itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs." R package version 2.3, https://CRAN.R-project.org/package=itsadug

Salvia, E., Bestelmeyer, P. E. G., Kotz, S. A., Rousselet, G. A., Pernet, C. R., Gross, J., & Belin, P. (2014). Single-subject analyses of magnetoencephalographic evoked responses to the acoustic properties of affective non-verbal vocalizations. Auditory Cognitive Neuroscience, 8, 422. https://doi.org/10.3389/fnins.2014.00422

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22, 1359–1366. https://doi.org/10.1177/0956797611417632

Singmann, H., Bolker, B., Westfall, J., & Aust, F., (2017). afex: Analysis of factorial experiments. R package version 0.18-0. https://CRAN.R-project.org/package=afex

Smith, N. J., & Kutas, M. (2015). Regression-based estimation of ERP waveforms: I. The rERP framework. Psychophysiology, 52, 157–168. https://doi.org/10.1111/psyp.12317

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage, 44, 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Soley-Bori, M. (2013). Dealing with missing data: Key assumptions and methods for applied analysis. Technical Report, 4, Boston University.

Strijkers, K., Costa, A., & Thierry, G. (2010). Tracking lexical access in speech production: Electrophysiological correlates of word frequency and cognate effects. Cerebral Cortex, 20, 912–928. https://doi.org/10.1093/cercor/bhp153

Toscano, J. C., McMurray, B., Dennhardt, J., & Luck, S. J. (2010). Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. Psychological Science, 21, 1532–1540. https://doi.org/10.1177/0956797610384142

ter Braak C. J . F. (1992). Permutation versus bootstrap significance tests in multiple regression and Anova. In K. H. Jökel, G. Rothe, & W. Sendler (Eds.), Bootstrapping and Related Techniques: Proceedings of an International Conference, Held in Trier, FRG, June 4–8, 1990, pp. 79–85. Springer-Verlag, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-48850-4 10

von der Malsburg, T., & Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language*, 94, 119-133. https://doi.org/10.1016/j.jml.2016.10.003

Vossen, H., Van Breukelen, G., Hermens, H., Van Os, J., & Lousberg, R. (2011). More potential in statistical analyses of event-related potentials: A mixed regression approach. International Journal of Methods in Psychiatric Research, 20, e56-68. https://doi.org/10.1002/mpr.348

Watkins, I. J., & Martire, K. A. (2015). Generalized linear mixed models for deception research: Avoiding problematic data aggregation. Psychology, Crime & Law, 21, 821–835. https://doi.org/10.1080/1068316X.2015.1054384

Westfall, J., Nichols, T. E., & Yarkoni, T. (2017). Fixing the stimulus-as-fixed-effect fallacy in task fMRI. Wellcome Open Research, *1*. https://doi.org/10.12688/wellcomeopenres.10298.2

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. NeuroImage, 92, 381–397. https://doi.org/10.1016/j.neuroimage.2014.01.060

Wood (2018). "mgcv" R package, https://CRAN.R-project.org/package=mgcv

Xiang, D. (2001). Fitting Generalized Additive Models with the GAM Procedure. SAS-SUGI Proceedings, Statistics, Data Analysis, and Data Mining, paper 256-26.

Table I. Monte-Carlo simulations of type I error rates when testing the main effects of factors A (between participants, within stimuli), B (within participants, between stimuli), and C (within participants, within stimuli) in a full factorial experiment with 20 participants and either 18 or 36 stimuli under the null hypothesis. Five methods are compared. Appropriate methods should return a proportion close to the nominal value of 0.05. Values significantly different from this value are in bold. The classical F1 approach has unacceptable type I error rates (up to more than 60%), suggesting that the use of this method fosters the replication crisis.

	18 stimulus			**36 stimulus**		
Method	**A**	**B**	**C**	**A**	**B**	**C**
Classical ANOVA -F1 (parametric)	**.0640 [.0568;.0721]**	**.6285 [.6137;.6436]**	**.1205 [.1108;.1310]**	.0555 [.0488;.0631]	**.5555 [.5403;.5711]**	**.0928 [.0842;.1022]**
Classical ANOVA-F1 (by permutation)	**.0638 [.0566;.0718]**	**.6258 [.6109;.6409]**	**.1202 [.1106;.1308]**	.0548 [.0481;.0623]	**.5542 [.5391;.5699]**	**.0930 [.0844;.1025]**
Mixed effect model	.0536 [.0458;.0626]	.0525 [.0448;.0615]	.0489 [.0415;.0576]	.0488 [.0425;.0559]	.0518 [.0453;.0591]	.0472 [.0411;.0543]
Quasi-F (parametric)	.0485 [.0423;.0556]	.0510 [.0446;.0583]	**.0362 [.0309;.0425]**	.0462 [.0402;.0532]	.0482 [.0420;.0554]	**.0380 [.0325;.0444]**
Quasi-F (by permutation, log p)	.0522 [.0458;.0596]	.0525 [.0460;.0599]	.0448 [.0388;.0516]	.0500 [.0437;.0572]	.0515 [.0451;.0588]	**.0430 [.0371;.0498]**

Table II. Monte-Carlo simulation of the proportion of FWER for the Quasi-F and F1 statistics when testing the main effect of factors A (between participants, within stimuli), B (within participants, between stimuli), and C (within participants, within stimuli) in a full factorial experiment with 20 participants and either 18 (top) or 36 (bottom) stimuli, using cluster mass tests and the Benjamini-Hochberg (BH) method to correct for multiple comparisons. Correct methods should be always close to the nominal value of 0.05.

Method	A	B	C
Classical ANOVA-F1 cluster	.1165 [.1070;.1269]	1 [1;1]	.2425 [.2296;.2561]
Classical ANOVA-F1 cluster (log p)	.1255 [.1156;.1362]	1 [1;1]	.2160 [.2036;.2291]
Classical ANOVA-F1 BH (parametric)	.0488 [.0425;.0559]	1 [1;1]	.2122 [.1999;.2253]
Quasi F cluster (log p)	.0815 [.0734;.0904]	.0660 [.0587;.0742]	.0910 [.0825;.1004]
Quasi F BH (parametric)	.0240 [.0197;.0293]	.0340 [.0288;.0401]	.0102 [.0075;.0139]

Method	A	B	C
Classical ANOVA-F1 Cluster	.0790 [.0711;.0878]	.9975 [.9959;.9990]	.1325 [.1224;.1434]
Classical ANOVA-F1 Cluster (log p)	.0828 [.0746;.0917]	.9970 [.9953;.9987]	.1218 [.1120;.1323]
Classical ANOVA-F1 BH (parametric)	.0385 [.0330;.0450]	1 [1;1]	.1205 [.1108;.1310]
Quasi F Cluster (log p)	.0693 [.0618;.0776]	.0792 [.0713;.0881]	.0910 [.0825;.1004]
Quasi F BH (parametric)	.0282 [.0235;.0339]	.0278 [.0231;.0333]	.0115 [.0086;.0153]

Figure captions

Figure 1. Figure 1. Re-analysis of a subset of picture naming data from Bürki (2017). A mixed-effects model was conducted at each time point, with the amplitude of the ERP as the dependent variable, condition (no versus neutral distractor) as the fixed-effect and random intercepts for participant and stimulus. The upper panel displays the standard deviation of the random intercept for participant and stimulus. The next two panels display the value of the random intercept for each participant and stimulus, respectively. The lower panel displays the fitted value of the fixed-effect. This example illustrates the need to take into account both the variability across participants and stimuli in the analysis.