Supplementary Material

In this appendix, we give an overview of some key statistical ideas that we did not have space to address in the main manuscript.

In section 1, we provide an introduction to two approaches to resampling statistics. In section 1.1, we give a broad overview of two such resampling approaches: permutation tests and bootstrapping methods. In section 1.2, we discuss the issue of constructing permutation tests for factorial ANOVA designs—the focus of the first aim of our simulation work.

In section 2, we provide an overview of the mass univariate corrections commonly used with EEG data. In section 2.1, we describe resampling-based corrections (F_{max} and cluster-based). In section 2.2, we describe corrections based on the false discovery rate.

1. Resampling statistics

The term resampling statistics refers to a broad set of non-parametric statistical methods that allow for inferences based on a sample of the population to the full population by resampling the observed data. Unlike parametric statistics, these methods do not assume any particular form of the probability distribution in the population. This means they can be used in cases where the assumptions of parametric tests are not valid. More generally, they can be used for test statistics for which there would otherwise be no way to ascertain the form of the probability distribution. This feature is key to their use in mass univariate corrections, as we discuss below (section 2.1).

1.1 Overview of two types of resampling approaches

The most common resampling approach for mass univariate statistics, and a focus of the present simulation work, are *permutation-based* approaches that randomly resample the data *without replacement*. That is, the data is re-assigned to condition labels, but each data point is used only once (for detailed discussion that goes beyond what we can provide here, see Good, 2005; Groppe, Urbach, & Kutas, 2011a; Manly, 2006; Maris & Oostenveld, 2007). For example,

consider a simple within-subjects priming experiment with a reaction time measure in the primed condition and a reaction time measure in the unprimed condition from each of 24 subjects. To conduct the permutation test, the primed and unprimed data are randomly swapped or not swapped for each subject independently. In other words, we create a situation in which the null is true by definition (because now the primed and unprimed conditions have data drawn from the same "population") but the distribution and variability of the data, including the subject-level blocking, is otherwise the same. A paired sample *t*-statistic is calculated for this randomly permuted data and this is repeated for a large number of random permutations.[1] The *t*-values across these permutations form the null distribution. The proportion of *t*-values in this null distribution that are as extreme or more extreme as the observed *t*-value from the unpermuted data is the *p*-value (i.e., the probability of obtaining a statistic as large as the observed *t* if the null hypothesis that the data in the two conditions were drawn from equivalent distributions were true).

When the assumptions of the parametric *t*-test are met, this empirical distribution of *t*-values will be a good approximation of the parametric *t*-distribution, and the permutation test will therefore give nearly the same answer as the parametric *t*-test. When the normality assumption of the *t*-test is not met, the permutation test will give a more accurate *p*-value than the parametric test. Thus, under conditions where both tests are justified, the parametric and permutation tests are approximately equally powerful and both maintain the Type I error rate appropriately, but the permutation test is justified in a wider set of circumstances.

An alternative resampling method (not considered in the present study) is based on bootstrap procedures. Unlike permutation tests, bootstrap procedures resample the data *with replacement*. This approach uses the observed data as an approximation of the population

[1] Ideally all possible permutations would be examined, but even with moderate sample sizes this is often impractical (e.g., with 24 subjects in a two-condition experiment, there are over 16 million permutations). Fortunately, a large number of random permutations will give a sufficiently precise estimate of the *p*-value. In fact, this is a straightforward binomial sampling problem, so the precision of the estimate can be calculated via a confidence interval.

distribution it was drawn from. We can then sample from this distribution repeatedly and calculate a relevant statistic from each sample to estimate the distribution of that statistic across repeated samples from the population. This allows for calculating a broader range of inferential statistics. For example, if one wants to calculate a confidence interval for the mean of a single group, it is possible to draw a large number of samples with replacement from the original sample and thereby construct a probability distribution of the mean. The 2.5 and 97.5 percentiles of this distribution give the limits of the bootstrap confidence interval.[2] For in depth discussion of bootstrap procedures, see Manly (2006), Good (2005), and Wilcox (2016).

1.2 The use of approximate permutation-based tests for factorial designs

As discussed in the main manuscript, for some effects in factorial designs, it is not possible to carry out permutation tests that control the Type I error rate at exactly the specified α. However, several methods of constructing an approximate test are possible. Here we briefly review the issue of permutation tests for factorial designs and the construction of exact and approximate tests.

For a one-way ANOVA, permutation-based versions of ANOVA using the *F*-statistic work exactly as described above for the *t*-test (i.e., the data for each subject is randomly reshuffled among all the conditions of the independent variable). However, consider a simple two-way design with factors A and B. Here, there are three effects in the ANOVA model: the main effect of A, the main effect of B, and the AxB interaction. In parametric ANOVA, the *F*-value calculated for each effect is compared to a different *F*-distribution depending on its degrees of freedom. Similarly, in the permutation ANOVA, separate null distributions must be constructed for each effect. The question that guides the permutations used to construct these null distributions is which data is exchangeable under the null hypothesis. For example, we can't test the effect of A

[2] This approach is called the "percentile bootstrap" and is given for illustrative purposes here because it is the simplest bootstrap approach in this case. However, it is important to note that it will not provide good coverage in all situations and more sophisticated bootstrap approaches are often preferable as discussed in the resources cited next.

by freely permuting data across all cells of the factorial design because the data is only exchangeable across all cells if the null is true for all three effects in the design; we wish to independently test whether the null hypothesis is true for A.

Appropriate exchangeability under the null can be achieved in three ways (discussed in greater detail in Anderson & Ter Braak, 2003; Welch, 1990). The first is via data reduction. For example, we can test the main effect of A by averaging across levels of B (within each subject and level of A), and then conducting a one-way permutation ANOVA on the reduced data (Welch, 1990). This achieves an exact test: that is, the Type I error rate is maintained at exactly the specified level. The second method is via restricted permutations. For example, to test the effect of factor A, we would only permute data across conditions of A while keeping each data point in the same condition of factor B. This provides an exact test only if we assume there is no AxB interaction (as the pattern of this interaction would not be held constant when the data is permuted). The third method is to subtract the effects that are not being tested—in this case the effect of B and AxB interaction—to obtain residuals such that the null effect is true for B and AxB. These residuals can then be permuted freely across all cells of the design under the assumption of the null hypothesis for A. However, the effects that are subtracted to form the residuals are only estimates of the true population effects. As a result, this permutation of residuals provides only an approximate test: the Type I error rate will be asymptotic to α as the sample size increases (because the estimate of the effects being subtracted is asymptotic to the population parameter as the sample size increases).

The main effect of A in the current example is a case where all three approaches are possible. In such cases, it is generally preferable to use a method that provides an exact test. However, for some interaction effects, exact tests are not possible. For the AxB interaction, if permutation is restricted within both A and B (and subjects), the only permutation possible is the original data, so the restricted permutation approach is not possible. In this case the method of data reduction is sometimes possible: if either factor A or B has only two levels, it is possible to

subtract across the levels of that factor and then conduct a one-way ANOVA (i.e., a test of whether the difference in factor A changes across levels of factor B is a test of the interaction effect). This will provide an exact test of the interaction. However, for designs in which more than one factor has more than two levels, only the permutation of residuals method is available.

The Factorial Mass Univariate Toolbox (FMUT) uses a combination of data reduction, restricted permutation, and permutation of residuals to construct a test that is as close to an exact test as possible. The exact approach is described in the FMUT documentation (https://github.com/ericcfields/FMUT/wiki).

2. Type 1 Error correction approaches for mass univariate analysis

The multiple comparison corrections used in mass univariate analysis generally fall into two types. One approach uses resampling approaches (as described above) to estimate the null distribution for a statistic that controls the probability of finding an effect across a set of time points and electrodes. The other uses probability theory to control the false discovery rate (FDR). Here, we provide a brief overview of these approaches. Detailed discussion can be found in Groppe et al. (2011a) and in citations provided for each correction below.

2.1. Resampling based corrections

There are two common mass univariate corrections that make use of resampling statistics. The first uses a resampling approach to estimate the distribution for the *maximum* effect across time points and electrodes assuming the null is true for all of them (Blair & Karniski, 1993). That is, for every permutation or bootstrapped sample of the data, the largest *t*-value or *F*-value across all time points and electrodes is recorded, and these form the distribution against which the observed statistics are compared. This correction, called t_{max} or F_{max}, is conceptually similar to a Bonferroni or Šidák correction. However, because the entire electrode x time point matrix is permuted or sampled for the bootstrap, it takes into account the correlations in the data across time points and electrodes. This provides a much less

conservative test than a Bonferroni or Šidák correction (which assume negatively correlated and independent data, respectively).

The second common correction makes use of cluster-based statistics. The most common version, the cluster mass test, finds adjacent time points/electrodes with effects surpassing some threshold (often the *t*-value or *F*-value that would be significant with no correction) and sums all the *t*s or *F*s to form a cluster mass statistic. A resampling approach is used to estimate the null distribution for this statistic by finding the largest cluster across a large number of random permutations or bootstrapped samples (Bullmore et al., 1999; Maris & Oostenveld, 2007). Other approaches to defining clusters have also been proposed and used, including threshold-free clustering techniques (Smith & Nichols, 2009). These approaches have been examined in simulation work with regard to Type I error rate (Pernet, Latinus, Nichols, & Rousselet, 2015), but it will be important for future work to compare various clustering approaches in terms of relative power to detect realistic ERP effects. All cluster approaches take advantage of the fact that true ERP effects tend to appear at several or many adjacent time points and electrodes, whereas much of the noise in the EEG data is more short-lived and/or confined to individual electrodes. Thus, large clusters are relatively rare in the null distribution, but are more likely for true effects.

In the present work, we focus on permutation-based versions of these corrections, rather than bootstrapped versions, because our goal was to extend the permutation-based methods that are already commonly used and that have received greater attention, use, and testing in the literature to date (e.g., Blair & Karniski, 1993; Groppe et al., 2011a; Groppe, Urbach, & Kutas, 2011b; Lage-Castellanos, Martinez-Montes, Hernandez-Cabrera, & Galan, 2010; Luck, 2014; Maris & Oostenveld, 2007). In addition, at least for Type I error rates, some previous simulation work suggests that permutation approaches are more accurate in small samples sizes (Pernet et al., 2015). However, bootstrapped versions of these corrections have been implemented in conjunction with two-step hierarchical linear modelling of EEG data (similar to standard MRI

analysis approaches) in the LIMO MEEG Toolbox (Pernet, Chauveau, Gaspar, & Rousselet, 2011), and it will be important for future work to explicitly examine the relative advantages of each approach via simulations that compare them using the same data and statistical designs.

2.2 False discovery rate corrections

Like the Bonferroni or Šidák corrections, FDR corrections are based on probability theory. Bonferroni and Šidák control the family-wise error rate: the probability that even one result in a group of tests will be a Type I error. False discovery rate corrections instead control the *proportion* of significant results that are Type I errors. Thus, if one uses a correction that controls the false discovery rate at 5% and finds 200 significant time points, then one should expect that around 10 of these are false positives. Whether this is a problem depends on whether one's conclusions depend on exactly which time points show a difference (in most ERP studies, we are only concerned that there is a difference in a general time window or on a particular component, and so this level of precision is unnecessary). There are several methods for calculating a false discovery rate correction depending on the assumptions that are made. The original Benjamini and Hochberg (1995) FDR correction assumes that the tests are either independent or positively correlated, which may not be valid for EEG data. Benjamini, Krieger, and Yekutieli (2006) developed a procedure that is intended to be less conservative when a large proportion of tests are testing a null effect (e.g., a short-lived component tested in a large time window), but which also assumes independent or positively correlated tests. Benjamini and Yekutieli (2001) introduced a formula to control the false discovery rate regardless of dependence, but which is also much more conservative (i.e., has reduced power).

In ERP analysis, FDR corrections are used by calculating a separate test at each time point and electrode, and then employing the correction to control the false discovery rate across all tests. This is generally employed with parametric *t*-tests or ANOVAs, but, as we note in the

main manuscript, one advantage of FDR corrections it that they can be combined with any model or statistical test, including non-parametric tests.

References

Anderson, M. J., & Ter Braak, C. J. F. (2003). Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation, 73*(2), 85-113. https://doi.org/10.1080/00949650210000155 58

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. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika, 93*(3), 491-507. https://doi.org/10.1093/biomet/93.3.491

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics, 29*(4), 1165-1188. https://doi.org/10.1214/aos/1013699998

Blair, R. C., & Karniski, W. (1993). An alternative method for significance testing of waveform difference potentials. *Psychophysiology, 30*(5), 518-524. https://doi.org/10.1111/j.1469-8986.1993.tb02075.x

Bullmore, E. T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., & Brammer, M. J. (1999). Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Transactions on Medical Imaging, 18*(1), 32-42. https://doi.org/10.1109/42.750253

Good, P. I. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses* (3rd ed.). New York, NY: Springer.

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011a). 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

Lage-Castellanos, A., Martinez-Montes, E., Hernandez-Cabrera, J. A., & Galan, L. (2010). False discovery rate and permutation test: An evaluation in ERP data analysis. *Statistics in Medicine, 29*(1), 63-74. https://doi.org/10.1002/sim.3784

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique* (2nd ed.). Cambridge, MA: The MIT Press.

Manly, B. F. J. (2006). *Randomization, Bootstrap and Monte Carlo Methods in Biology* (3rd ed.). London, UK: Chapman & Hall.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177-190. https://doi.org/10.1016/j.jneumeth.2007.03.024

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*(831409). 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*, 85-93. https://doi.org/10.1016/j.jneumeth.2014.08.003

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage, 44*(1), 83-98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Welch, W. J. (1990). Construction of permutation tests. *Journal of the American Statistical Association, 85*(411), 693-698. https://doi.org/10.2307/2290004

Wilcox, R. R. (2016). *Introduction to Robust Estimation and Hypothesis Testing* (4th ed.). Waltham, MA: Elsevier.