

Randomization-based Hypothesis Testing from Event-related Data

Richard E. Greenblatt, Ph.D. and Mark E. Pflieger, Ph.D.

Source Signal Imaging, San Diego CA 92102 USA

Running title: Hypothesis testing

Corresponding author: Richard E. Greenblatt, Ph.D.
Source Signal Imaging, Inc.
2323 Broadway #102
San Diego CA 92102 USA

Abstract

Methods are described for non-parametric significance testing from event-related encephalographic data, using randomization tests. These methods may be applied in both signal space and source space. The methods include within-subject between-condition comparisons, paired and unpaired comparisons, and within-group and between-group comparisons. Test statistics are also derived for comparing the spatial or temporal response patterns, independent of specific changes at individual locations. Novel methods for testing peak-height significance, and also for making map-wide comparisons, are described. These methods have been validated using simulated data.

Keywords: Hypothesis testing, event related potential, permutation test, statistics, human electrophysiology, nonparametric methods, statistical nonparametric mapping

hypothesis (Karniski *et al.*, 1994; Holmes *et al.*, 1996; Nichols and Holmes, 2001; Raz *et al.*, 2003). The great advantage of these methods is that they permit us to estimate distribution functions and covariance structures from the data themselves, without requiring *a priori* assumptions, such as Gaussianity, required by parametric methods (*e.g.*, Friston *et al.*, 1995). However, in cases where the distribution is Gaussian, parametric and non-parametric methods will obtain essentially equivalent results.

First, we describe these methods at a fairly general level, and then show how they may be applied to widely-used evoked response experimental designs. While the use of randomization tests is not new when applied to ERP data (*e.g.*, Karniski *et al.*, 1994), we propose methods based on a classification scheme that accounts for a broader class of experimental analyses than have been used previously, to the best of our knowledge. We also describe a novel method for estimating the significance of ERP peaks, the “plus-minus” test.

Methods

Cell-based observation statistics.¹ Given $M \in \mathbb{N}$ observation locations, $N \in \mathbb{N}$ events (trials or subjects), and a trial interval T with latency $\tau \in \mathbb{Z}$. Let $\mathbf{V}_{n,c} \in \mathbb{R}^{M \times T}$ be the matrix representation for the n^{th} event and the c^{th} condition. Let $\mathbf{V}_{n,c}[m, \tau]$ represent the observation at location m and latency τ for the n^{th} event and the c^{th} condition. We

¹ A note on notation: Scalars are denoted by lower case italics (*e.g.*, s), vectors are denoted by lower case bold text (*e.g.*, \mathbf{v}), maps and matrices (including row and column vectors, when considered as matrices) by upper case bold (*e.g.*, \mathbf{T}). \mathbb{N} represents the set of integers greater than 0, \mathbb{Z} , the set of integers, and \mathbb{R} , the set of real numbers.

sometimes refer to each element of the matrix $\mathbf{V}_{n,c}$ as a *cell*, and the entire matrix as a *map*. In this case, the map is equivalent to a scalar field over a discrete lattice. However, we wish to generalize the concept of a map to include fields of other types (such as vector fields) over the lattice. The lattice points are typically parameterized by space and time, although other parameterizations, such as frequency or time-frequency, may also be assimilated readily into the concepts we develop.

Note that the physical interpretation of the map cells may vary (e.g., the voxel intensities of an image, measured voltages at electrode locations, current source density estimates on the cortical surface, etc.), but these may all be treated in a uniform manner.

Vector field maps. When the map is a vector field, some additional complications arise. All cases begin with a collection of vector-valued maps $\{\bar{\mathbf{M}}_i\}$ (scalar-valued maps are included as a special case). It is assumed that the maps have been normalized for multiple comparisons, as described below. The maps are sorted into two groups, A and B. Randomization of group assignment (denoted by $*\{\cdot\}$) may occur, which may be any permutation (for unpaired tests) or a permutation constrained to associated pairs (for paired tests). Observation statistic maps (denoted by \mathbf{T}) are computed for each grouping. The observation statistic may be the mean, standard deviation, variance, root mean square, power, or some other (possibly higher order) statistic. In cases 2-4 below, vector-valued maps are converted to scalar-valued maps by computing the norm, or vector magnitude, in each cell. (For scalar-valued maps, magnitude = absolute value.) Map differences are performed either between vector-valued or magnitude maps to produce delta maps.

Case 1 may be represented by the following diagram:

$$*\{\vec{M}_i\} \begin{array}{l} \xrightarrow{A} \vec{T}_A \\ \xrightarrow{B} \vec{T}_B \end{array} \xrightarrow{\vec{T}_A - \vec{T}_B} \vec{\Delta}$$

The key feature of case 1 is that maps are vector-valued throughout, without ever converting to map magnitudes. Thus, the number of obtained p -values equals the product of the number of cells and the number of vector space dimensions. The delta maps are signed. Case 1 includes the important special case of one-tailed tests for scalar-valued maps (for both paired and unpaired randomizations).

Case 2 has the following diagram:

$$*\{\vec{M}_i\} \begin{array}{l} \xrightarrow{A} \vec{T}_A \\ \xrightarrow{B} \vec{T}_B \end{array} \xrightarrow{\vec{T}_A - \vec{T}_B} \vec{\Delta} \xrightarrow{|\vec{\Delta}|} \Delta$$

Case 2 is the same as case 1 with the additional step of converting to magnitude maps at the end. The resultant delta maps are non-negative. This case covers two-tailed tests for scalar input maps.

Case 3 has the following diagram:

$$*\{\vec{M}_i\} \begin{array}{l} \xrightarrow{A} \vec{T}_A \xrightarrow{\|\vec{T}_A\|} \mathbf{T}_A \\ \xrightarrow{B} \vec{T}_B \xrightarrow{\|\vec{T}_B\|} \mathbf{T}_B \end{array} \xrightarrow{\mathbf{T}_A - \mathbf{T}_B} \Delta$$

Compared with case 2, note that map magnitudes in case 3 are obtained before the map difference. Thus, the delta maps are signed. Case 2 includes the “phase” (temporal and/or spatial) in the difference, whereas case 3 ignores this “phase” and pays attention only to magnitude differences.

Case 4 has the following diagram:

$$\{\vec{\mathbf{M}}_i\} \xrightarrow{\|\vec{\mathbf{M}}_i\|} * \{\mathbf{M}_i\} \begin{array}{l} \xrightarrow{A} \mathbf{T}_A \\ \xrightarrow{B} \mathbf{T}_B \end{array} \xrightarrow{\mathbf{T}_A - \mathbf{T}_B} \Delta$$

Here the map magnitudes are computed prior to grouping. Since case 4 reduced the vector-field map to a scalar field map prior to permutation, from the hypothesis testing standpoint, it is equivalent to a scalar field map test.

A classification scheme for hypothesis testing with event-related data. In Figure 1, we illustrate a dichotomous scheme for the classification of event-related experimental data. The classification depends on number of comparisons, and subject vs. group comparisons, as defined in the figure, but is independent of preprocessing steps. In particular, it generalizes to both signal and source spaces. Although randomization-based ANOVA tests can be assimilated into this scheme, we will not consider this problem here.

Insert Figure 1 near here.

Paired comparisons. A *paired comparison* typically arises when we want to test the effect of 2 different conditions on a response. For example, we may have measured an evoked response to some stimulus before and after training within a group of N subjects. Another example of a paired comparison is one in which we wish to compare the first to the second instance of a physically equivalent stimulus.

Consider the within-group, between-condition case, where we have 2 conditions, c_0 and c_1 . For each subject n , location m and latency τ , we compute the difference between the statistics for the 2 conditions

$$\Delta_{(c_1-c_0)} = \mathbf{T}_{c_1} [m, \tau] - \mathbf{T}_{c_0} [m, \tau] \quad (1.1)$$

Then the null hypothesis is that there is no difference between the 2 conditions, or

$$H_0 : \Delta_{(c_1-c_0)} = \mathbf{0} \quad (1.2)$$

where $\mathbf{0}$ is a (column) vector of zeroes.

This null hypothesis is equivalent to assuming that the test statistics \mathbf{T}_{n,c_0} and \mathbf{T}_{n,c_1} differ only by labeling. Thus if the null hypothesis is true, we should be able to change the order of the terms randomly in (1.1) without changing the expected value of $\Delta_{(c_1-c_0)}$.

In virtue of the pairing, there must be the same number of observations (or subjects) for each of the 2 conditions. This is represented diagrammatically in (1.3) as 2 parallel columns, where each column represents the (ordered) test statistic matrix set $\{\mathbf{T}_c\}$.

$$\begin{bmatrix} \mathbf{T}_{0,c_0} \\ \mathbf{T}_{1,c_0} \\ \equiv \\ \mathbf{T}_{N-1,c_0} \end{bmatrix} \begin{bmatrix} \mathbf{T}_{0,c_1} \\ \mathbf{T}_{1,c_1} \\ \equiv \\ \mathbf{T}_{N-1,c_1} \end{bmatrix} \quad (1.3)$$

A randomization based on (1.3) is equivalent to exchanging the labeling (column ordering) for random rows. Since any such relabelling can be represented by an N-bit binary number (e.g. take 0 to represent the original column order for a selected row, and 1 the column-reversed order), it follows that the number of combinations for the paired case is

$$NC_{\text{paired}} = 2^N \quad (1.4)$$

Unpaired comparisons. An *unpaired comparison* may arise when, for example, we wish to compare 2 conditions (say rare and frequent stimuli) within a single subject, or within a group of subjects. In this case, the order of the individual events in computing the test statistic is not relevant, and we may have differing numbers of observations, N_0 and N_1 of the 2 conditions. Then

$$H_0 : \Delta_{(c_1-c_0)} = \mathbf{0} \tag{1.5}$$

The null hypothesis is formally the same as that in the paired case. The null hypothesis is again equivalent to the assertion that the observations differ only in their labeling, so that label exchanges should not change the expected value of $\Delta_{(c_1-c_0)}$, but now the ordering of observations is ignored, consistent with (1.5). This may be represented diagrammatically by a single column in (1.6)

$$\left[\begin{array}{c} \mathbf{T}_{0,c_0} \\ \mathbf{T}_{1,c_0} \\ \equiv \\ \mathbf{T}_{N_0-1,c_0} \\ \dots\dots\dots \\ \mathbf{T}_{0,c_1} \\ \mathbf{T}_{1,c_1} \\ \equiv \\ \mathbf{T}_{N_1-1,c_1} \end{array} \right] \tag{1.6}$$

The entries above and below the dotted line in (1.6) represent the terms on the right and left of the difference in (1.5), respectively. A random relabelling is equivalent to swapping some number of observation statistics across the dotted line. If we assume that $N_1 \geq N_0$, this can be achieved by randomly selecting (without replacement) N_1 times from the set represented in (1.6), and assigning these to the first term in (1.5), then

assigning the remaining N_0 terms to the second term. Equivalently, the only ordering that matters is whether an observation shows up above or below the dotted line, as long as the number of terms above and below does not change. From this description, the number of unpaired combinations is given by

$$NC_{\text{unpaired}} = \frac{(N_0 + N_1)!}{(N_0)!(N_1)!} \quad (1.7)$$

Multiple comparison correction. Nichols and Holmes (2001) present a method for multiple comparison correction in medical imaging data, based on the work of Westfall and Young (1993). These methods can be readily adapted to evoked response statistical maps, in particular the *single threshold test*. We compute the maximal test statistic for each permutation, where the maximal test statistic is defined as the largest value of the test statistic realization over the entire space-time map. This gives a permutation distribution, \mathbf{T}^{Max} , for the map-wide maximal test statistic. Then for a selected p -value, α , the critical threshold for \mathbf{T}^{Max} (or \mathbf{T}^{Min}) is given by the $c+1$ largest value, where c is given by $\lfloor \alpha J \rfloor$, αJ rounded down, where J is the number of permutations used. “Voxels with statistics exceeding this threshold exhibit evidence against the corresponding voxel hypotheses at level α . The corresponding corrected p -value for each voxel is the proportion of the permutation distribution for the maximal statistic that is greater or equal to the voxel statistic.” (Nichols and Holmes, 2002).

The single threshold test is a conservative multiple comparison correction, and controls the family-wise error rate (Benjamini and Hochberg, 1995), similar to a Bonferroni correction.

Normalization for multiple correction correction. When the map is a scalar field, one commonly used observation statistic is the mean, given by

$$\mathbf{T}_\mu = \bar{\mathbf{V}} = \frac{1}{N_c} \sum_{N_c} \mathbf{V}_n \quad (1.8)$$

where N_c ranges over all trials of the appropriate condition.

Although it is possible to carry through the randomization tests using the mean as a test statistic, its use adds some additional complexity when dealing with the multiple comparison problem. For this reason, it is often desirable to use a standardized test statistic, such as the t -statistic, defined in (1.9) for the cell $[m, \tau]$ as

$$t[m, \tau] = \frac{1}{N} \sum_N \frac{v_n[m, \tau] - \bar{v}[m, \tau]}{\sqrt{s^2[m, \tau]}/N} \quad (1.9)$$

where the variance of cell $[m, \tau]$, $s^2[m, \tau]$, is given by

$$s^2[m, \tau] = \frac{1}{N-1} \sum_N (v_n[m, \tau] - \bar{v}[m, \tau])^2 \quad (1.10)$$

Note that in equations (1.8)-(1.10), N ranges over all maps in both conditions.

Estimating the variance of vector field maps begins by computing the covariance matrix at each cell. In Case 1, where vector values are maintained throughout, the diagonal entries may be used to normalize the individual vector component amplitudes. Alternatively the vector space may be orthogonalized. In cases 2-4, where the vectors are converted to a scalar field prior to hypothesis testing, the covariance must be reduced to a variance. We can do this by summing the covariance eigenvalues.

More formally, let $\mathbf{W}_n[m, \tau] \equiv \mathbf{V}_n[m, \tau] - \bar{\mathbf{V}}_n[m, \tau]$. Then

$$\mathbf{C}[m, \tau] \equiv \frac{1}{N-1} \sum_N \mathbf{W}_n[m, \tau] \mathbf{W}_n^T[m, \tau] \quad (1.11)$$

is the covariance of cell $[m, \tau]$.

Let $\lambda_d[m, \tau]$ be the d^{th} eigenvalues of $\mathbf{C}[m, \tau]$. Then, if D is the within-cell vector space dimension,

$$s^2[m, \tau] = \sum_D \lambda_d[m, \tau] \quad (1.12)$$

is the variance of cell $[m, \tau]$.

Special case: signal vs. background. Often experimenters would like to know if a phase-locked peak (say the auditory N100) is significantly different from noise. While it is possible to compare the peak amplitude to an estimate of background (e.g., pre-stimulus interval), this assumes that the variation is stationary with respect to the event of interest, and this stationarity assumption is violated more and more as the temporal distance between the event of interest and the “background” interval increases. Fortunately, this problem can be addressed by means of a specialized randomization test, which we call the “plus-minus” (or “ \pm ”) test, for reasons that will become apparent shortly.

In most cases, this problem is equivalent to asking if the mean value for the m^{th} cell (e.g., electrode) at time sample t , $\bar{v}_{m,t}$, is significantly different from zero. Then the null hypothesis is given by

$$H_0 : \bar{v}_{m,t} = 0 \quad (1.13)$$

To test the null hypothesis, we need to make one additional assumption, that the variation of the null distribution is symmetric around 0 (note that the distribution need not be Gaussian, however). Random realizations of this distribution may be obtained from the experimental data by selecting single trials, multiplying each selection by either +1 or -1 with equal probability (hence the name “plus-minus”), then summing to obtain a realization. This method is derived from the work of Schimmel (1967). Note that the covariance structure is also unaffected by the (\pm) randomization (Pflieger *et al.*, 1995).

If the null hypothesis were true, the expected value of this sum is clearly 0, and the variance estimates the variance of the hypothetical distribution at that time point. Then, as in a conventional permutation test, we can rank the realized values at each time point and map location, and compare the true value to the realized distribution to obtain a p -value.

Vector marginal contrasts. Considering a scalar field map as a matrix, $\mathbf{V}_{n,c}$, the map can be viewed as a collection of row and column vectors, and these vectors may take on meaningful physical interpretations. For example, the M -dimensional column vector at latency τ , $\mathbf{V}_{n,c}[:, \tau]$, represents the spatial pattern of the observations at time τ , typically a vector in signal or source space. Similarly, the T -dimensional row vector $\mathbf{V}_{n,c}[m, \cdot]$ represents the temporal pattern observed at location m . This vectorization is useful for developing statistical tests that are sensitive to spatial or temporal patterns in the data, but do not depend directly on differences at individual locations.

We can make use of these vectors to develop some statistical tests that are sensitive to these patterns.

Given 2 vectors, say $\mathbf{V}_{n,c_0} [\cdot, \tau]$ and $\mathbf{V}_{n,c_1} [\cdot, \tau]$, corresponding to the spatial pattern vectors under conditions c_0 and c_1 , respectively, there are three natural comparisons between these vectors that will result in a single number each, the observation statistics that form the basis for a randomization test. These are

(i) the distance between two vectors in the appropriate normed vector space,

$$\|\Delta\|_{c_1-c_0} = \|\mathbf{V}_{n,c_1} [\cdot, \tau] - \mathbf{V}_{n,c_0} [\cdot, \tau]\| \quad (1.14)$$

(ii) the correlation between the two vectors

$$R_{c_1,c_0} = \frac{\mathbf{V}_{n,c_1} [\cdot, \tau] \cdot \mathbf{V}_{n,c_0} [\cdot, \tau]}{\|\mathbf{V}_{n,c_1} [\cdot, \tau]\| \|\mathbf{V}_{n,c_0} [\cdot, \tau]\|} \quad (1.15)$$

(iii) the difference in magnitude between the two vectors

$$\Delta_{\|\cdot\|_{c_1} - \|\cdot\|_{c_0}} = \|\mathbf{V}_{n,c_1} [\cdot, \tau]\| - \|\mathbf{V}_{n,c_0} [\cdot, \tau]\| \quad (1.16)$$

Note that (1.15) measures pattern differences independent of power, while (1.16) measures power differences independent of pattern, while (1.14) combines both effects. Each of these observation statistics (1.14)-(1.16) may be associated with its null hypothesis

$$H_0 : \|\Delta\|_{c_1-c_0} = 0 \quad (1.17)$$

$$H_0 : R_{c_1,c_0} = 1 \quad (1.18)$$

$$H_0 : \Delta_{\|\cdot\|_{c_1} - \|\cdot\|_{c_0}} = 0 \quad (1.19)$$

Having formulated the appropriate null hypotheses, essentially that same randomization tests may be applied as those that we described for testing contrasts between conditions within individual cells.

Results

In order to verify the methods, and also to evaluate the resulting software, we have carried a series of experiments using simulated data. Here we will consider the verification of the 1-condition (plus-minus) test and the 2-condition, within-subject, unpaired comparison in signal space. These two cases permit a test of the principal assumptions on which the randomization methods are based. In order to verify the methods, we need a “gold standard”, so we have chosen to generate simulated scalp data with a simple statistical structure so that the results can be compared directly with well-known parametric tests. When the statistical properties of the data are known (Gaussian in this case), parametric and non-parametric methods are expected to provide equivalent results. However, the great advantage of randomization methods is that they will estimate the statistical structure from the data themselves, and therefore will continue to be valid without assuming that the underlying distribution is known *a priori*.

Simulated data. Data were simulated using EMSE v5.0 software (Source Signal Imaging, Inc., San Diego CA). A single radial dipole was placed eccentrically in a 0.1m radius 3-shell sphere, and the forward solution was computed for 47 locations on the sphere surface. Pseudorandom, independent, identically distributed, Gaussian noise was added linearly to each of the 47 channels. 100 trials were simulated using a common average reference, each consisting of a single square wave, as shown in Figure 2. A

second set of 100 trials was generated using statistically equivalent but non-identical noise, but without the dipole, for use in the 2-condition test.

Insert Figure 2 near here.

± test. A topographic p -value map was obtained for the simulated data, using the (\pm) test as implemented in EMSE v5.0, as shown in Figure 3(A) for a representative time point when the dipolar square wave was active. 100 randomizations were used, allowing a p -value resolution of 0.01. The p -values are two-tailed (*i.e.*, significant positive excursions have p -values near to 1.0, and significant negative excursions have p -values near to 0.0).

To compare the (\pm) test results with conventional parametric test results, we first computed the average and variance across all trials. The latency-by-channel variance was then used to convert the average data to Z -values, from which the latency-by-channel p -values could be obtained from the Gaussian cumulative density function. The resulting topography for a representative latency is shown in Figure 3(B). Clearly, the results from the randomization and parametric tests are essentially equivalent.

Insert Figure 3 near here.

Within subject between conditions, unpaired comparison. A topographic p -value map was obtained for the simulated data, using the within-subject between condition test as implemented in EMSE v5.0. Condition 1 employed the same set of 100

trials with a square wave dipolar source that was used in the (\pm) test. Condition 2 used a time series with statistically equivalent Gaussian noise but without any added dipolar signal. 100 permutations were used, allowing a p -value resolution of 0.01. The resulting 1-tailed p -map is shown in Figure 3(C) for a representative time point when the dipolar square wave was active, with significant p -values near to 0.0.

To compare the permutation test with a conventional parametric test, we used the independent samples t -test with Bonferroni correction for multiple comparisons (Altman, 1991). The results are shown in Figure 3(C). The results from the randomization and parametric tests are essentially similar, though there are some differences in regions of marginal significance. These differences may be attributable to the random sampling inherent in the randomization tests.

Discussion

The utility of significance testing with event-related data is clear. Both parametric (*e.g.*, Duffy, 1981) and non-parametric (Karniski *et al.*, 1994) methods have been applied to scalp data. Parametric methods have been proposed for electrophysiological source space data (Greenblatt and Gao, 1999), but these methods suffer from significant difficulties associated with required assumptions about the statistical structure of the data, especially with regard to the multiple comparison problem. Randomization tests provide a way out of this problem, since the statistical structure is inferred directly from the data. In this work, we propose a unified framework for hypothesis testing in both signal space and source space, using randomization methods. Pascual-Marqui (unpublished) has implemented similar methods for the Loreta

source estimation method (Pascual-Marqui *et al.*, 1995), a special case of the methods we describe.

In order to apply these methods, the user must first identify the observation statistic and null hypothesis to be tested. As described in the Methods section, the data are then randomly relabeled (in the two condition test) to obtain the cell-based distribution. The true observation statistic difference may be compared to this distribution to obtain a p -value, indicating that a difference this large or larger might be due to chance, under the null hypothesis. The plus-minus test proceeds in an analogous fashion.

Experiments with simulated data that we describe have demonstrated the validity of this approach. These methods are currently available in EMSE v5.0 software, and are being evaluated with human-subject data.

Literature Cited

Altman, DG. Practical Statistics for Medical Research. Chapman and Hall, London, 1991.

Benjamini Y and Hochberg Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. B.*, 1995, **57(1)**:289-300.

Duffy, FH, Bartels PH and Burchfiel JL. Significance Probability Mapping: An Aid in the Topographic Analysis of Brain Electrical Activity. *EEG Clin. Neurophysiol.* 1981, 51:455-462.

Friston KJ, Holmes AP, Worsley KJ, Poline J-P, Frith CD, and Frackowiak RSJ. Statistical Parametric Maps in Functional Imaging: A General Linear Approach. *Human Brain Mapping*, 1995 **2**:189-210.

Greenblatt, RE and Gao, L.. M/EEG Source Space Statistical Parametric Mapping. *NeuroImage*, 1999, **9**:S156.

Holmes AP, Blair RC, Watson JDG, Ford I. Non-Parametric Analysis of Statistic Images From Functional Mapping Experiments. *Journal of Cerebral Blood Flow and Metabolism*, 1996, **16**:7-22.

Karniski W, Blair RC, and Snider AD. An Exact Statistical Method for Comparing Topographic Maps, with Any Number of Subjects and Electrodes. *Brain Topography*, 1994, **6(3)**:203-210.

Nichols TE and Holmes AP. Nonparametric Analysis of PET Functional Neuroimaging Experiments: A Primer. *Human Brain Mapping*, 2001, **15**:1-25..

Pascual-Marqui RD, Michel CM, and Lehmann D. Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. *Intl. J. Psychophysiol.*,1994, **18**:49-65.

Pflieger ME, Simpson GV, Vaughn HG. Improved estimation of ERP source activities in the presence of realistic background EEG. *Human Brain Mapping*, 1995, **Suppl. 1**:101.

Raz J, Zheng H, Ombao H, Turetsky B. Statistical tests for fMRI based on experimental randomization. *NeuroImage*, 2003, **19**:226-32.

Schimmel, H. The (\pm) reference: Accuracy of estimated mean components in average response studies. *Science*, 1967, **157**:92-94.

Westfall PH and Young SS. *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*. Wiley, New York, 1993.

Acknowledgements

This work was supported in part by grant NS36133, National Institute of Neurological Diseases and Stroke (U.S.), to REG.

Figure Legends

Figure 1. We employ a dichotomous classification scheme for hypothesis testing, which can be represented by analogy with a cladogram, as illustrated in the figure. Each node of the tree represented a dichotomous choice (*e.g.*, between one condition vs. > 1 condition) and the terminal leaves represent the test description (*e.g.*, 2-condition, paired within-subject comparison). The ANOVA test is shown connected with a dotted line, because we do not consider >2 condition tests in this paper.

Figure 2. Single trial simulated data for three representative channels are shown in the upper row. The corresponding 100-trial averages for the same channels are shown in the lower row. These simulated data were used for both 1-condition and 2-condition tests, as described in the text.

Figure 3. Topographic p-value maps were obtained for 1 condition vs. baseline (A and B), and 2-condition (C and D), using randomization methods (A and C) and parametric methods (B and D), as described in the text. Contour intervals are at increments of $p=0.05$, and the small black dots indicate the electrode locations.

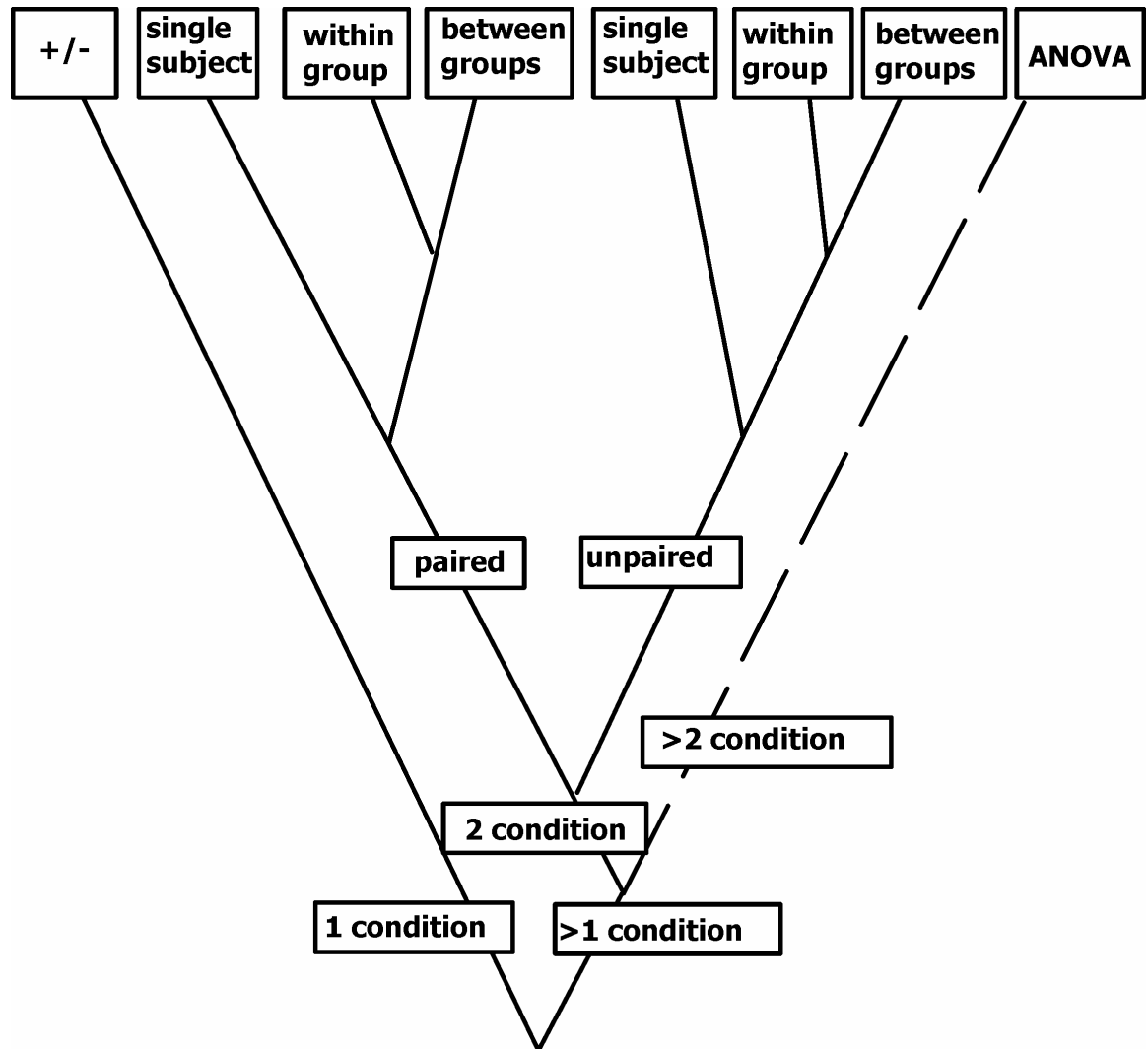


Figure 1.

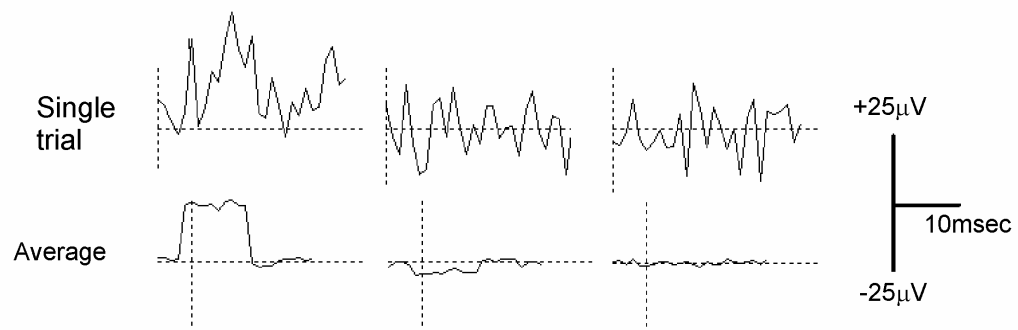


Figure 2.

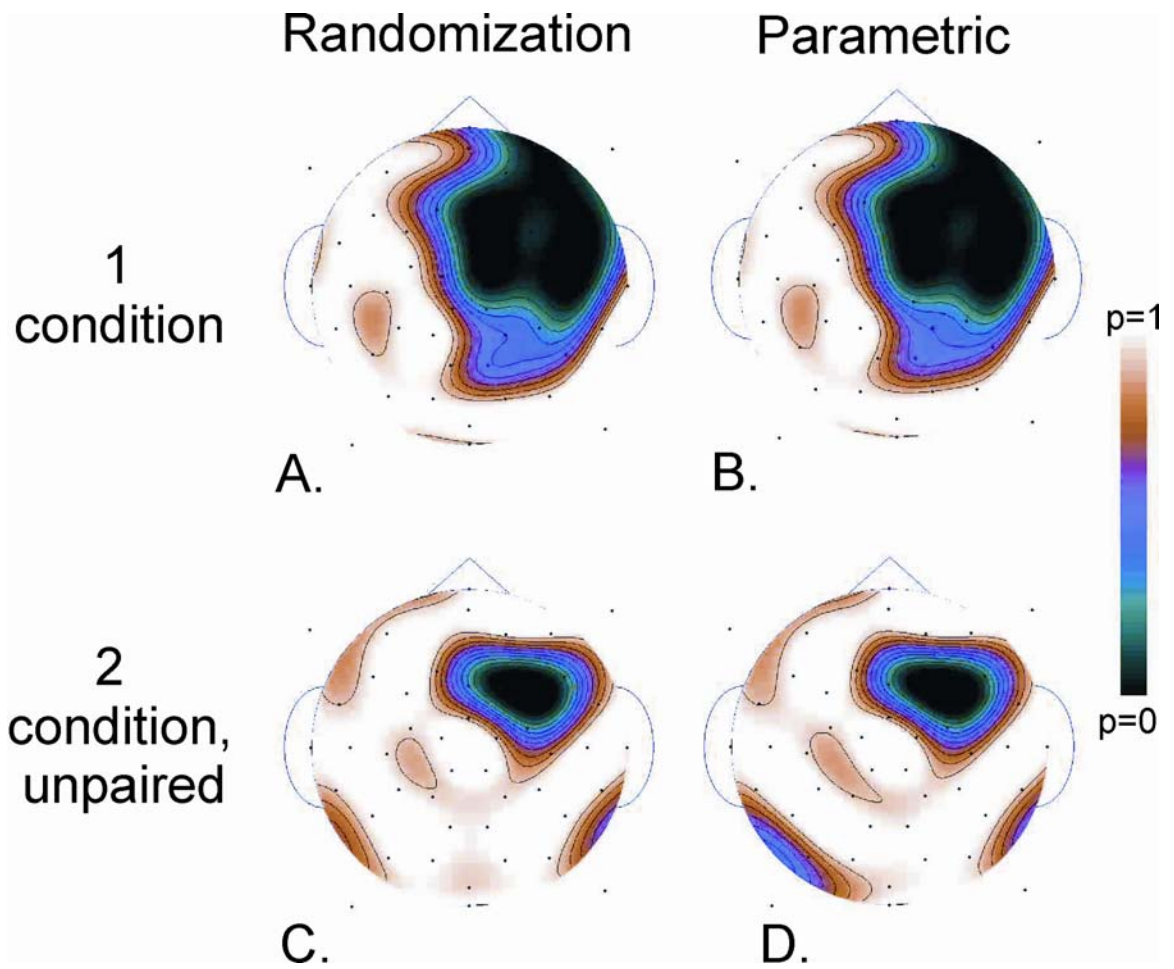


Figure 3.