*Psychophysiology, 41* (2004), 142–151. Blackwell Publishing Inc. Printed in the USA.
© 2003 Society for Psychophysiological Research
DOI: 10.1111/j.1469-8986.2003.00139.x

# Randomization tests for ERP topographies and whole spatiotemporal data matrices

ERIC MARIS
Nijmegen Institute of Cognition and Information, University of Nijmegen, 6500 HE Nijmegen, The Netherlands

**Abstract**

In ERP studies, the comparison of topographies (multichannel measurements) or whole spatiotemporal data matrices (multichannel time series of measurements), the classical statistical tests very often cannot be used. It is argued that, for these comparisons, *randomization tests* are an excellent alternative. It is also argued that the randomization test is superior to another resampling method, the *bootstrap*, because exact probability statements (e.g., *p* values) can be made. A review is given of the literature on randomization tests designed for electrophysiological data. New randomization tests are presented and applied to two data sets, one coming from a psychopharmacological experiment and the other from an ERP experiment in visual word recognition.

**Descriptors:** Randomization tests, Family-wise error rate, Electrophysiological data, Topographies, Spatiotemporal data

In this article, I present a class of statistical tests of differences between event-related potentials (ERPs) among a number of conditions. The focus is on the statistical comparison of quantities in cases where the classical statistical tests very often cannot be used: topographies (i.e., multichannel measurements) and whole spatiotemporal data matrices (i.e., multichannel time series of measurements). It is argued that, for these comparisons, *randomization tests* are an excellent alternative and better than a related method, the *bootstrap*. For illustration purposes, randomization tests were applied to two data sets, one coming from a psychopharmacological experiment and the other from an ERP experiment in visual word recognition.

## Setup

I consider multichannel ERP data observed over a number of time points, as determined by the time interval of measurement and the sampling rate of the registration equipment (usually, between 200 and 1000 Hz). This type of data is called spatiotemporal and can be organized in a matrix of order number of channels × number of time points.

I consider ERPs observed in a number of conditions. These conditions may differ because, for example, different types of stimuli were presented, different instructions were given, or the participants differed in a systematic way (e.g., patients vs. controls). In every condition, a number of replications is observed. In single-participant studies, these replications are the single-trial ERPs; in multiple-participant studies, these replications are the ERPs of a number of participants. Typically, in multiple-participant studies, the ERPs of individual participants are themselves averages of a number of single-trial ERPs. In this article, a distinction is made between two types of multiple-participant studies: studies with a between-participants manipulation, in which every participant is observed in only one of the conditions, and studies with a within-participants manipulation, in which every participant is observed in all conditions. This distinction is made because different statistics are used in between- and within-participant studies.

In the following, I first deal with the statistical testing of the difference between topographies. Later, I will consider the statistical testing of the difference between whole spatiotemporal data matrices. Both statistical testing problems will be considered for within- as well as between-participant studies.

## Statistical Testing of the Difference between Topographies

A topography is a multichannel measurement at a particular latency[1] (e.g., at 400 ms). Statistical testing of the difference between topographies is of interest for two reasons:

Thanks to Dick Willems for his assistence in this project, Marijtje Jongsma for the use of the data of the first application, Inge Keus for her assistence in the collection of the data of the second application, and Chan Wai Yen for proofreading the manuscript.

Address reprint requests to: Eric Maris, Nijmegen Institute of Cognition and Information (NICI), University of Nijmegen, P.O. Box 9104, 6500 HE Nijmegen, The Netherlands. E-mail: maris@nici.kun.nl.

[1] In practice, very often, instead of considering a measurement at a particular latency, the average is taken over multichannel measurements in a given time window (e.g., between 370 and 430 ms poststimulus). This averaging has no implications for the statistical test, and therefore I will ignore the difference between time-averaged and nontime-averaged measurements.

1. Topographies reflect the sources that generate the ERPs. In other words, the topography of an ERP is considered a *scalp signature* of the source that generated it. Therefore, by comparing topographies of different conditions, it is possible to obtain estimates of the underlying sources that are responsible for the differences between these conditions.
2. Testing the difference between all channels (i.e., topographies) by means of a single statistic, as is proposed in this article, controls the family-wise error rate[2] without any correction of the Bonferroni type. This approach is valid for researchers who are not interested in the topography of the effect of an independent variable but only in whether this effect is statistically significant. The simplest method to evaluate the statistical significance of an effect is by performing a series of channel-specific statistical tests. As is well known, this leads to an uncontrolled increase in the family-wise error (FWE) rate, which can be bounded by Bonferroni correction, but most likely at the expense of an excessive decrease in power.

The statistic that is computed to perform the statistical test of the difference between the conditions depends on the number of conditions that are compared. However, the main points of this article do not depend on the number of conditions that are compared, and therefore these points will be elaborated for the case of two conditions. After that, it is easy to generalize to the case of an arbitrary number of conditions.

In the next section, I present the multivariate independent samples *T* test for the difference between independent samples, performed by means of Hotelling's $T^2$ statistic. This test is used for testing the difference between two conditions in (a) a single-participant study (in which the replications are single-trial ERPs),[3] and (b) a multiple-participant study with a between-participants manipulation (in which the replications are the ERPs of a number of participants). In a multiple-participant study with a within-participants manipulation, one has to use the multivariate independent samples *T* test for the difference between *paired* samples, which is performed by means of another version of Hotelling's $T^2$ statistic. I will return to this test for paired samples in one of the later sections.

### *The Multivariate T Test for the Difference between Two Independent Samples*

The multivariate independent samples *T* test is a generalization of the univariate independent samples *T* test. The number of replications in the first and the second samples is denoted by, respectively, $n_1$ and $n_2$. The replications themselves are denoted by, respectively, $y_{1i}$ ($i = 1,\ldots,n_1$) and $y_{2i}$ (i = $1,\ldots,n_2$). Every $y_{1i}$ and $y_{2i}$ denotes a multichannel measurement observed at a particular latency or in a particular time window. Thus, $y_{1i}$ and $y_{2i}$ are vectors of length equal to the number of channels. In the following, the number of channels is denoted by $M$. The average measurements in the two samples are denoted by $\bar{y}_1$and $\bar{y}_2$.

The multivariate independent samples *T* test makes use of Hotelling's $T^2$ statistic:

$$T^2 = (\bar{y}_1 - \bar{y}_2)^t \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{\text{pooled}}\right]^{-1} (\bar{y}_1 - \bar{y}_2). \quad (1)$$

To fully understand the right-hand side of this equation, one has to know some elementary matrix algebra. However, for the purpose of this article, it is sufficient to have an intuitive understanding of this formula. In a loose sense, $T^2$ can be seen as the square of the vector $\bar{y}_1 - \bar{y}_2$ divided by the matrix $\frac{1}{n_1} + \frac{1}{n_2}$ $S_{\text{pooled}}$ (which explains the exponent $^{-1}$ in Equation 1). The value of $T^2$ is always positive. A good handbook on applied multivariate statistics (e.g., Johnson & Wichern, 1988) provides information on the properties and applications of Hotelling's $T^2$.

The vector $\bar{y}_1 - \bar{y}_2$ is the difference between the average multichannel measurements in the two conditions. The more the elements of $\bar{y}_1 - \bar{y}_2$ differ from zero, the larger $T^2$. The symbol $S_{\text{pooled}}$ denotes the pooled variance–covariance matrix of the multichannel measurements. *Pooled* means that, in $S_{\text{pooled}}$, the variance–covariance matrices of both conditions are combined by taking their weighted average (with the weights proportional to the sample sizes). The diagonal elements of $S_{\text{pooled}}$ are the pooled variances of the different channels, and the off-diagonal elements are the pooled covariances between the different pairs of channels. The larger the pooled variances on the diagonal, the smaller $T^2$. The role of the pooled covariances is more complicated. The pooled variance–covariance matrix is multiplied by $\frac{1}{n_1} + \frac{1}{n_2}$. The larger $n_1$ and $n_2$, the number of observations in the two samples, the larger $T^2$.

It is important to note here that the $T^2$ statistic cannot be computed with average reference data. This is because, with average reference data, the inverse of the variance–covariance matrix $S_{\text{pooled}}$ does not exist. With a common reference electrode, either physical or synthetic (e.g., linked mastoids), this inverse exists if the number of observations exceeds a certain critical number (see further).

The $T^2$ statistic in Equation 1 is a generalization of the square of the univariate (single-channel) independent samples *T* statistic. To see this, consider the $T^2$ statistic in Equation 1 for the case of a single channel. In this case, $\bar{y}_1 - \bar{y}_2$ is a single number, as is $S_{\text{pooled}}$, the pooled variance of the measurements in this single channel. Let this pooled variance be denoted by $s^2_{\text{pooled}}$. Then, the $T^2$ statistic can be written as follows:

$$T^2 = \frac{(\bar{y}_1 - \bar{y}_2)^2}{(\frac{1}{n_1} + \frac{1}{n_2}) s^2_{\text{pooled}}}, \quad (2)$$

which is the square of the univariate independent samples *T* statistic.

Under the null hypothesis that the expected values of $y_{1i}$ and $y_{2i}$ are equal, and some auxiliary assumptions (see further), the distribution of the $T^2$ statistic is a scaled $\mathscr{F}$ distribution with $M$ and $n_1 + n_2 - M - 1$ degrees of freedom. (By *scaling*, I mean multiplication by a known constant. For the remainder of this article, it is not necessary to know the value of this scaling constant.) By evaluating the observed $T^2$ statistic under this reference distribution, the FWE rate can be controlled. The auxiliary assumptions that also have to be fulfilled (besides the null hypothesis) for the $T^2$ statistic to have the usual reference

[2] The family-wise error rate is the probability under the null hypothesis of observing a significant effect at one or more channels.

[3] There is some confusion in the psychophysiological community as to whether the independent samples *T* test or the paired samples *T* test should be used in a single-participant study. Whether the one or the other statistical test should be used depends on whether the data come in pairs, with one element of the pair observed in one condition and the other element in the other condition. Only if the data come in pairs should the paired samples *T* test be used. This is because it is likely that the elements of a pair are correlated. Contrary to a multiple-participant study with a within-participants manipulation, there are no such pairs in a single-participant study.

distribution given above, are the following:

1. The vectors $y_{1i}$ and $y_{2i}$ are drawn from multivariate normal distributions with identical variance–covariance matrices.
2. The total number of observations (i.e., $n_1+n_2$) is larger than the number of channels plus one (i.e., $M+1$). If this condition does not hold, then the matrix inverse in Equation 1 does not exist and therefore the $T^2$ statistic cannot be computed.

Although both auxiliary assumptions can be violated, in the psychophysiological literature, the latter assumption has received more attention.

### *Many Channels and Few Observations*

In many ERP studies, the total number of observations is less than the number of channels plus one, and therefore Hotelling's $T^2$ statistic cannot be computed. In the literature, several proposals have been made for dealing with this problem (Achim, 2001; Galán, Biscay, Rodríguez, Pérez-Abalo, & Rodríguez, 1997; Haig & Gordon, 1995; Karnisky, Blair, & Snider, 1994). All proposals, in one way or another, reduce the multivariate testing problem to a uni- or a bivariate problem. For instance, Karnisky et al. proposed to compute a separate $T$ statistic for every channel, take the square of these statistics, and add these values over the channels.[4] Their statistic is called $T^2=\text{sum}$. Because the sampling distribution of $T^2=\text{sum}$ is unknown, Kawrnisky et al. used the randomization distribution of this statistic to control the FWE rate. The randomization distribution will be described in the next section.

Haig and Gordon (1995) proposed to project the topographies onto the so-called *centroid difference vector*, and to compute the signed projection lengths. This centroid difference vector is the difference between the centroids, the average topographies in the two samples. The signed[5] projection lengths are a measure of relative distance of the topographies to the two centroids. To make this more explicit, let $x_1$ and $x_2$ be topographies, belonging to, respectively, sample 1 and sample 2. Also, let the centroid difference vector be computed by subtracting the average topography (centroid) of sample 2 from the average topography (centroid) of sample 1. Then, if there is a difference between the samples, the topography $x_1$ is expected to be closer to the centroid of sample 1 than to the centroid of sample 2, and for the topography $x_2$ the reverse is expected. This will reflect itself in a larger signed projection length when $x_1$ is projected onto the centroid difference vector than when $x_2$ is projected onto this vector. Completely in line with this fact, Haig and Gordon propose to perform an ordinary independent samples $T$ test on these projections. However, Achim (2001) noted that the assumptions of the independent samples $T$ test were not fulfilled by the procedure of Haig and Gordon, and that therefore the usual reference distribution (i.e., the $\mathscr{T}$-distribution) is not the sampling distribution of this test statistic. Achim proposed to use the randomization distribution as an alternative reference distribution instead.

Several other test statistics have been proposed, and I briefly mention them here. Galán et al. (1997) proposed to compare topographies by first locating, in each of the conditions, the electrodes with maximum potential, and then computing the angle between these two scalp locations.[6] Galán used the randomization distribution to control the FWE rate. Finally, Achim (2001) proposed two test statistics that are based on a principal components analysis of the data. In essence, he proposed to perform a $T$ test on the first or the first two principal components (in the latter case, by means of the bivariate $T^2$ statistic).

### *Randomization Tests for Independent Samples*

In this article, a randomization test is proposed for ERP data. A randomization test involves the following two steps: (a) Compute some statistic in which you are interested, and (b) evaluate its value under the randomization distribution. First, I will discuss the randomization distribution, and especially, the motivation for its use. And second, I will discuss the choice of the test statistic whose value will be evaluated under the randomization distribution. For concreteness, assume that the test statistic is the $T^2$ statistic for independent samples in Equation 1. Instead of evaluating this test statistic under its usual reference distribution (a scaled $\mathscr{F}$-distribution), it will now be evaluated under the randomization distribution.

*The randomization distribution.* The randomization distribution arises as a result of random assignment of units to conditions. The randomization distribution can be used in many experimental designs, but for the purpose of introduction, it is good to consider a multiple-participant study with a between-participants manipulation. In a study of this type, the units are the participants and these are assigned at random to one of the conditions.

To understand the rationale behind the use of the randomization distribution, one should start from the hypothesis of no effect. This hypothesis of no effect has a very precise definition: The value that is observed for a particular participant is identical to the value that would be observed if this participant were assigned to the other condition. Under this hypothesis, if the participants were assigned differently to the conditions, the same $n_1+n_2$ values would be observed but distributed differently over the two conditions. The random element in these data is the way the observations are distributed over the two conditions. This random element defines a probability distribution over the data, which is called the randomization distribution.

The randomization distribution of any statistic of the data can be simulated on a computer. Suppose that the randomization mechanism that performed the actual assignment was such that every partition of the $n_1+n_2$ participants in two groups of sizes $n_1$ and $n_2$ were equally likely. Such a randomization mechanism can easily be simulated on a computer by making use of a pseudorandom number generator, which is available in almost all programming languages. The randomization distribution of, say, the $T^2$ statistic can be approximated by means of an algorithm that repeats the following three steps a large number of times (e.g., 10,000 times):

[4] Karnisky et al. (1994) considered a within-participants comparison of conditions, and therefore computed paired samples $T$ statistics, but their method can easily be extended to a between-participants comparison, namely, by computing independent samples $T$ statistics.

[5] Signed means that the direction of the projection is also taken into account. More specifically, if the signed projection length is positive, this means that the topography points more in the direction of the centroid difference vector than in the opposite direction. And if the signed projection length is negative, the reverse holds.

[6] This test was originally proposed for a within-participants comparison, but it is easy to formulate a variant for a between-participants comparison.

1. Perform a random partition of the $n_1+n_2$ observed values in two groups such that every possible partition is equally likely.
2. Compute the $T^2$ statistic for this partition.
3. Add the value of this statistic to a temporary list of $T^2$ statistics computed under the hypothesis of no effect.

The final list of $T^2$ statistics is an estimate of the randomization distribution of this statistic, and it can be used to estimate both some quantile (e.g., the 95th) of the randomization distribution as well as the $p$ value of the observed $T^2$ statistic. This $p$ value is simply the proportion of $T^2$ statistics under the hypothesis of no effect that is larger than the observed $T^2$ statistic.

Now, consider a single-participant study. In a study of this type, the units are the occasions at which the trials can be presented. Typically, these occasions are time-slots in an experimental session. These occasions are assigned at random to one of the two conditions. Under the assumption of no effect, the value that is observed at a particular occasion is identical to the value that would be observed if this occasion was assigned to the other condition. Under this hypothesis, if the occasions were assigned differently to the conditions, the same $n_1+n_2$ values would be observed but distributed differently over the two conditions. Because the assignment is random, it defines a probability distribution over the data. This probability distribution can be simulated in the same way as described above.

The hypothesis of no effect will also be called a *null hypothesis*, but it should be noted that this null hypothesis is different from the null hypothesis of the multivariate independent samples $T$ test, which states that the expected values of the multichannel measurements in the two conditions, $y_{1i}$ and $y_{2i}$, are equal. In fact, the hypothesis of no effect is a hypothesis about the $n_1+n_2$ units that were observed (namely, that their values are identical in the two conditions), whereas the null hypothesis of the multivariate independent samples $T$ test is about the expected values in the populations from which the samples were drawn.[7] There seems to be no reason to prefer one null hypothesis over the other; they are just different specifications of the general idea of no difference between the conditions.

*The choice of the test statistic.* For every test statistic, it is possible to approximate its randomization distribution by means of the simulation algorithm described above. Thus, by means of the randomization distribution, it is possible to control the FWE rate for the test statistics proposed by Achim (2001), Galán et al. (1997), Haig and Gordon (1995), and Karnisky et al. (1994). This obviates the need to derive their sampling distribution, a job that may be very difficult or even impossible.

The univariate statistics proposed by the authors above solve the problem that, in many ERP studies, the total number of observations is larger than the number of channels plus one, such that the $T^2$ statistic cannot be computed. However, many other test statistics can be conceived for the same purpose. I now propose one such statistic that stays much closer to Hotelling's $T^2$ than the univariate statistics proposed by other authors. The main advantage of using Hotelling's $T^2$ is that it is equivalent to using the so-called *likelihood ratio statistic* (Johnson & Wichern, 1988), which has certain optimum properties with respect to power. As to its computation, contrary to the univariate statistics, Hotelling's $T^2$ takes into account the covariances between pairs of channels, which are always nonzero in practice.[8]

The new statistic to be presented is based on the idea of pooling the variance–covariance matrices $S_{\text{pooled}}$ over adjacent time points. This is also called a *moving average* of time-specific variance–covariance matrices and it results in a *double-pooled* variance–covariance matrix, denoted by $S_{\text{double-pooled}}$. The single-pooled variance–covariance matrix $S_{\text{pooled}}$ in the formula of Hotelling's $T^2$ is then replaced by this double-pooled variance–covariance matrix $S_{\text{double-pooled}}$.

To make this more precise, let the variance–covariance matrix of time point $t$ be denoted by $S_{\text{pooled}}$. If the number of observations is less than the number of channels plus one, this matrix cannot be inverted (i.e., the $^{-1}$ operation cannot be performed). However, consider now the pooling of $S_{\text{pooled}}$ over the adjacent time points $t-1$, $t$, and $t+1$, resulting in the double-pooled variance–covariance matrix:

$$S_{\text{double-pooled}} = \frac{S^{(t-1)}_{\text{pooled}} + S^{(t)}_{\text{pooled}} + S^{(t+1)}_{\text{pooled}}}{3},$$

which is a moving average of width three. By increasing the number of adjacent time points that is involved in this moving average (3, 4, 5, ...), one can always compute a double-pooled variance–covariance matrix that is inverible.

If the single-pooled $S_{\text{pooled}}$ in the formula of Hotelling's $T^2$ is replaced by $S_{\text{double-pooled}}$, the usual reference distribution for the $T^2$ statistic (a scaled $\mathscr{F}$-distribution) is no longer valid. However, the FWE rate can also be controlled by using the randomization instead of this sampling distribution.

*The randomization test solves several problems.* Not only can the randomization distribution be determined for every test statistic, this possibility does not depend on any other auxiliairy assumption besides the fact that the statistic must be computable. This is different from the usual reference distributions, which are all sampling distributions under the null hypothesis *plus* some nontrivial auxiliary assumptions. For instance, the usual reference distribution for Hotelling's $T^2$ (a scaled $\mathscr{F}$-distribution) is only valid if the observations are drawn from multivariate normal distributions with an identical variance–covariance matrix in the two conditions. If there is reason to doubt these auxiliary assumptions, it is wise to consider the randomization distribution as an alternative reference distribution.

*Post hoc testing of separate channels.* When a significant Hotelling's $T^2$ is observed, the obvious next question is which channels are responsible for the effect. In other words, given a significant difference between the two conditions, the question is where this difference occurs. This question can be answered by means of a randomization test that is based on a combination of squared univariate (channel-specific) $T$ statistics. This test resembles Scheffé's post hoc test, which is often used in a common univariate ANOVA following a significant omnibus test (see Maxwell & Dalaney, 1990, pp. 186–192). The test statistic for this randomization test is computed as follows:

1. For each of the $M$ channels, compute the squared univariate $T$ statistic in Equation 2.
2. Compute the maximum of the $M$ squared univariate $T$ statistics.

[7] If the null hypothesis of the randomization test holds, then, under random assignment to the conditions, also the null hypothesis in terms of expected values (namely, $\mathscr{E}(Y_1) = \mathscr{E}(Y_2)$) holds. The reverse does not necessarily hold.

[8] If these covariances are all zero, then Hotelling's $T^2$ is equal to $T^2 =$ sum, proposed by Karnisky et al. (1994).

From a large number of draws from the randomization distribution of the maximum of the squared univariate $T$ statistics, estimate the 95th quantile of the randomization distribution. Denoting the number of draws by $D$, this estimate is equal to the .95 × $D$ largest value in the list of $D$ draws. This estimated 95th quantile is then used as a critical value to which all squared univariate $T$ statistics are compared: For every channel with a squared univariate $T$ statistic that is larger than this critical value, it is concluded that it exhibits a significant difference between the conditions.

This post hoc testing procedure controls the FWE rate for all channels jointly. This is because, under the randomization null hypothesis of no effect, there is a probability of .05 of having one or more channels with a squared univariate $T$ statistic that is larger than the 95th quantile of the randomization distribution of the maximum of all squared univariate $T$ statistics.

### *Comparing More than Two Conditions*

Hotelling's $T^2$ is restricted to cases where only two conditions have to be compared. When more than two conditions have to be compared, the usual test statistic is Wilks' lambda.[9] Wilks' lambda is the multivariate generalization of the ANOVA $F$ statistic. For two independent samples, Hotelling's $T^2$ (the multivariate generalization of the squared $T$ statistic) is a simple function of Wilks' lambda and results in exactly the same $p$ value.

If the number of observations (summed over all conditions) is small as compared to the number of channels (e.g., for three groups, smaller than or equal to the number of channels minus two), then Wilks' lambda cannot be used. Again, this problem can be solved by replacing the time-specific single-pooled variance–covariance matrix (which is also needed for the computation of Wilks' lambda) by a double-pooled variance–covariance matrix. For evaluating the significance of this adaptation of Wilks' lambda, one cannot make use of a distributional result that gives us its sampling distribution. However, by making use of the randomization distribution, it is easy to control the FWE rate. Post hoc testing of separate channels is possible in essentially the same way as was described for the case of exactly two conditions: Instead of taking the maximum of squared univariate (channel-specific) $T$ statistics, one now has take the maximum of channel-specific $F$ statistics.

### *The Multivariate $T$ Test for the Difference between Paired Samples*

I now consider multichannel potentials observed in two experimental conditions of a factor that is manipulated within participants. The observations, which come in $n$ pairs, are again denoted by $y_{1i}$ and $y_{2i}$. The multivariate paired samples $T$ test makes use of Hotelling's $T^2$ statistic:

$$T^2 = (\bar{y}_1 - \bar{y}_2)^t \left[\frac{1}{n} S_{\text{diff}}\right]^{-1} (\bar{y}_1 - \bar{y}_2). \tag{3}$$

The symbol $S_{\text{diff}}$ denotes the variance–covariance matrix of the differences $y_{1i} - y_{2i}$. The diagonal elements of $S_{\text{diff}}$ are the variances of the differences between the potentials at the different channels, and the off-diagonal elements are the covariances between these differences at the different pairs of channels.

Much of what was said for the multivariate independent samples $T$ test can be repeated here for the multivariate paired samples $T$ test with minor modifications. First, the $T^2$ statistic in Equation 3 is a generalization of the square of the univariate (single-channel) paired samples $T$ statistic.

Second, under the null hypothesis that the expected values of $y_{1i}$ and $y_{2i}$ are equal, and some auxiliary assumptions (see further), the distribution of the $T^2$ statistic is distributed as a scaled $\mathscr{F}$-distribution with $M$ and $n - M$ degrees of freedom. The auxiliary assumptions are the following:

1. The difference vectors $y_{1i} - y_{2i}$ are drawn from a multivariate normal distribution.
2. The number of pairs $n$ is larger than the number of channels $M$. If this condition does not hold, then $S_{\text{diff}}$ is not invertible and therefore $T^2$ cannot be computed.

Third, the idea of pooling the variance–covariance matrices of adjacent time points can also be applied to paired samples. More specifically, the matrix $S_{\text{diff}}$ in Equation 3 can be replaced by another matrix that is obtained by pooling the $S_{\text{diff}}$ matrices observed in some small time interval. Because its sampling distribution is unknown, the resulting test statistic should be evaluated under its randomization distribution.

Fourth, the randomization distribution for studies with a within-participants manipulation arises from a slightly different random process as for studies with a between-participants manipulation. In studies with a between-participants manipulation, the randomization distribution arises from the random assignment of participants to conditions, whereas in studies with a within-participants manipulation, the randomization distribution arises from random assignment of *occasions* to conditions. Typically, these occasions are time slots in an experimental session, and one trial is presented in every time slot. In an experiment with two conditions, the time slots are divided in two sets, and one set is assigned at random to one condition and the other set to the other condition. The measurements on the trials belonging to each of the two conditions are then averaged over the trials. In the above, these averages are denoted by $y_{1i}$ and $y_{2i}$.

The assumption of no effect involves that exactly the same potentials would be observed if the time slots were assigned in the reverse way: The time slots that were first assigned to one condition are now assigned to the other. The result of this reversal is that the values of $y_{1i}$ and $y_{2i}$ are interchanged and that the difference $y_{1i} - y_{2i}$ changes sign. Because the assignment of occasions to conditions is random, it defines a probability distribution over the data. In particular, under the hypothesis of no effect, the probability of observing the difference $y_{1i} - y_{2i}$ is equal to the probability of observing minus this difference.

Fifth, the randomization distribution of any statistic of the data (i.e., the $n$ differences $y_{1i} - y_{2i}$) can be simulated on a computer. The randomization distribution of, say, the paired samples $T^2$ statistic with a pooled variance–covariance matrix can be approximated by means of an algorithm that repeats the following three steps a large number of times:

1. For every participant, simulate the toss of a fair coin (with probability .5 of observing heads) and let the result of this toss determine whether the observed difference $y_{1i} - y_{2i}$ is multiplied by +1 or −1.
2. Compute the $T^2$ statistic using the differences computed in step 1.

[9] Together with Wilks' lambda, statistical packages often report Pillai's trace, Roy's largest root, and Hotelling's trace. (The latter statistic is not Hotelling's $T^2$ statistic.) These four multivariate test statistics very often lead to the same conclusion, and usually only Wilks' lambda is reported in an article.

3. Add the value of this $T^2$ statistic to a temporary list of $T^2$ statistics computed under the hypothesis of no effect.

The final list of $T^2$ statistics is an estimate of the randomization distribution of this statistic, and it can be used to estimate both some quantile of the randomization distribution as well as the $p$ value of the observed $T^2$ statistic.

Sixth, post hoc testing of separate channels is performed in essentially the same way as for a between-subjects manipulation with two levels: Instead of taking the maximum of squared univariate (channel-specific) independent samples $T$ statistics, one now has to take the maximum of squared univariate (channel-specific) paired samples $T$ statistics. For the rest, the post hoc testing procedure is identical.

Seventh and last, if the within-participants manipulation has more than two levels, a slighty different version of the $T^2$ statistic has to be used. For instance, if the manipulation has three levels, then two differences have to be computed for every participant: for instance, $y_{1i} - y_{2i}$ and $y_{1i} - y_{3i}$. These two difference vectors of length $M$ are then concatenated, producing a combined difference vector of length $2 \times M$. In the computation of the test statistic, this combined difference vector replaces the single difference vector $y_{1i} - y_{2i}$ in our discussion of the two-level case. The generalization to the case of an arbitrary number of levels is along the same lines as for the case of three levels.

## Statistical Testing of the Difference between Multichannel Time Series of Measurements (Spatiotemporal Data)

We now consider the statistical testing of the difference between multichannel time series of measurements. A multichannel time series of measurements is organized in a $M \times S$ spatiotemporal data matrix (with $M$ and $S$ denoting, respectively, the number of channels and the number of time points). The observed spatiotemporal data matrices in the two conditions are denoted by, respectively, $y_{1i}$ and $y_{2i}$.

### *A Straightforward Statistical Test that Is Infeasible in Practice*

A straightforward statistical test for comparing the spatiotemporal data matrices of two conditions is a $M \times S$-variate $T^2$ statistic. In the previous section, the interest was in comparing topographies, which are multichannel measurements observed at a single time point, consisting of $M$ elements. Now, complete spatiotemporal matrices, consisting of $M \times S$ elements, are considered. Therefore, for comparing the spatiotemporal data matrices of two conditions, a $M \times S$-variate $T^2$ statistic is required. To be able to compute this $M \times S$-variate $T^2$ statistic, the number of observations (participants or trials) has to be larger than $M \times S+1$. Because the number of channels $M$ is usually between 10 and 100, and the number of time points $S$ is usually of the order of several hunderds, an extremely large number of observations is required. It is very unlikely that such a large number of observations is attainable in practice.

Besides being infeasible in practice, this statistical test also has the disadvantage that, if it is significant, it gives no information about the time points at which the difference between the conditions occurs. In the following, a statistical test will be presented that does give this information.

### *A Feasible Randomization Test*

We now consider a randomization test that is based on a combination of test statistics for several time points. This test resembles the post hoc testing procedure that was proposed for the statistical testing of the difference between topographies. The test statistic is computed as follows:

1. For every time point from 1 to $S$, compute the appropriate $T^2$ statistic: In the case of two independent samples, compute the $T^2$ statistic in Equation 1, and in the case of paired samples, compute the $T^2$ statistic in Equation 3.
2. Compute the maximum of the $S$ $T^2$ statistics. This combined statistic will be denoted as Max($T^2$).

From a large number of draws from the randomization distribution of Max($T^2$), the 95th quantile of this distribution is estimated. This estimated quantile is then used as a critical value to which all time-point-specific $T^2$ statistics are compared: For every time point with a $T^2$ statistic that is larger than this critical value, it is concluded that it exhibits a significant difference between the conditions.

This procedure controls the FWE rate for all time points jointly. This is because, under the randomization null hypothesis of no effect, there is a probability of .05 of having one or more time points with a $T^2$ statistic that is larger than the 95th quantile of its randomization distribution.

It is instructive to see that the FWE rate for all time points jointly is not controlled when the randomization test of the previous section (i.e., for a single time point) is applied to all $S$ time points separately. To see this, suppose that the randomization null hypothesis of no effect holds and that the researcher evaluates all $S$ $T^2$ statistics under their own (time-point-specific) randomization distribution. Let the significance level be .05. Then, it can be shown that the expected number of time points with a significant $T^2$ statistic is equal to $S \times .05$. Thus, when $S$ is equal to 1,000, one can expect 50 time points with a significant $T^2$ statistic. It is clear that, in this case, the FWE rate for all time points jointly is not controlled at .05.

It is worth considering a couple of variations on the Max($T^2$) randomization test. First, it may happen that the $T^2$ statistics cannot be computed because the number of observations is too small. In that case, one can use the same solution as described in the previous section: taking the moving average of the time-specific variance–covariance matrices. If there is reason to believe that the variance–covariance matrices are constant over time, then it makes sense to take the average of all $S$ time-specific variance–covariance matrices. This will result in a more reliable estimate of the variance–covariance matrix and therefore in a more sensitive statistical test (at least, if the variance–covariance matrices are indeed constant over time). However, in my own experience with data from cognitive neuroscience, the variance–covariance matrices are not constant over time; usually, the variances increase over time.

Second, it may happen that a researcher is not interested in finding significant differences at a single time point only. This is likely to be true if the sampling rate is very high (e.g., 500 Hz or more). For researchers who are not interested in effects that last only 1 or 2 ms, it makes sense to compute a moving average on the time series of $T^2$ statistics, with the width of the moving average equal to the minimum time interval that is large enough to be of interest. If the Max($T^2$) randomization test is applied to a time series of moving averages instead of to the time series of the original $T^2$ statistics, a single moving average that exceeds the critical value under its randomization distribution now corresponds to an interval instead of a single time point. An advantage of taking a moving average of the time series of $T^2$ statistics is that, if the effect lasts as long as the width of the moving average, this will increase the power of the statistical test. This is because an average $T^2$ statistic is more reliable than individual $T^2$ statistics.

The widths of the moving averages of the time-specific variance–covariance matrices and $T^2$ statistics are two tuning parameters of the randomization test that is presented here. It should be clear that these tuning parameters should be fixed in advance, because the type 1 error probability may not be controlled otherwise. In the absence of theoretical results on the effects of these tuning parameters, the best option is to determine the tuning parameters from an application of the randomization test to independent pilot data. In the applications discussed in the next section, the widths of the two moving averages are fixed at 31 and 10 ms for, respectively, the time-specific variance–covariance matrices and the $T^2$ statistics.

Third, it is easy to extend this test of the difference between two spatiotemporal data matrices to the case of more than two conditions. For the case of more than two independent samples, the Max($T^2$) statistic is replaced by the maximum of $S$ Wilks' lambdas, each computed at a single time point. And for the case of a within-participants manipulation with more than two levels, the Max($T^2$) statistic is replaced by the maximum of $S$ time-specific $T^2$ statistics, each of which is computed on a *combined* difference vector, as described previously.

## A Comparison with Other Resampling Methods

The randomization test belongs to the larger class of resampling methods. All resampling methods attempt to quantify uncertainty by resampling from the set of observations. A randomization test is a resampling method because it can be conceived as a procedure in which (1) all observations in the different conditions are put into one or more urns, and (2) new values for the observations in the conditions are found by drawing at random *without* replacement from these urns. By repeating steps 1 and 2, draws from the randomization distribution are obtained.

Two other well-known resampling methods are the *bootstrap* and the *jackknife*. Both methods are described by Wasserman and Bockenholt (1989) in a paper that was especially written for the psychophysiological community. Wasserman and Bockenholt focused on the estimation of parameters (a population mean, a correlation, a regression coefficient, etc.) and in particular, on how to quantify the uncertainty in these estimates. For example, Fabiani, Gratton, Corballis, Cheng, and Friedman (1998) applied the bootstrap method to quantify the uncertainty in the location (Cz, Pz, F7, etc.) that has the largest P3 amplitude. In this study, the parameter is the electrode placement at which the largest P3 amplitude is observed, and the uncertainty in the parameter estimate is quantified by a frequency distribution over the 30 placements used in this study; if the frequencies are strongly concentrated around a single placement, this means that there is not much uncertainty about the locus of the largest P3 amplitude.

This focus on parameter estimation differs from the present article, in which the focus is on the testing of the hypothesis of no effect. With unidimensional measurements (e.g., a measurement at a particular sensor at a particular time point), it is very often possible to reformulate the hypothesis of no effect as a hypothesis about the value of some unknown parameter (e.g., a population mean), and this usually leads to a simple statistical test. However, when the hypothesis is about multidimensional measurements (i.e., topographies and whole spatiotemporal data matrices), this reformulation is much more complicated and, more importantly, does not lead to a simple statistical test.

There is an easy bootstrap method for testing the hypothesis of no effect with multidimensional measurements, but this method is not along the lines described by Wasserman and Bockenholt (1989). This method will be described in the following. The jackknife, in contrast, cannot readily be used for testing this hypothesis and will be ignored in the following.

I now briefly describe how the bootstrap method can be used to test the hypothesis of no effect. The focus is on the comparison of topographies in a within-participants design with two conditions. The extension to whole spatiotemporal data matrices, independent samples (a between-participants design or a single-participant study), and more than two conditions is simple. For the application to the comparison of topographies in a within-participants design with two conditions, the bootstrap method involves the following four steps:

1. Compute *residuals*. For the type of data that is considered here, these residuals are the observed difference topographies $y_{1i} - y_{2i}$ *minus* their average $\bar{y}_1 - \bar{y}_2$.
2. Put all $n$ residuals in a single urn.
3. Draw $n$ residuals at random *with replacement* from this urn.
4. Compute Hotelling's $T^2$ statistic in Equation 3, treating the residuals as the actual observations $y_{1i} - y_{2i}$.

By repeating steps 3 and 4 a large number of times, the so-called *bootstrap distribution* of $T^2$ statistics is obtained. By evaluating the observed $T^2$ statistic under this bootstrap distribution (comparing the observed $T^2$ statistic with some quantile of this distribution or computing a $p$ value), the statistical significance of this observed $T^2$ statistic is assessed.

The rationale of the bootstrap method is that the bootstrap distribution is considered as an approximation of the sampling distribution under the null hypothesis. In fact, if the number of observations $n$ is infinite, then the bootstrap distribution is identical to this sampling distribution. One should not be completely reassured by this result, because statistical testing only makes sense in finite samples; in infinite samples, there is no uncertainty about whether or not the null hypothesis holds.

The main difference between the randomization test and the bootstrap is that the randomization test is exact: The randomization test makes an exact probability statement (e.g., a $p$ value) about the observations under the hypothesis of no effect. This probability statement follows from the randomization mechanism that is responsible for the assignment of participants or occasions to conditions. In contrast, a probability statement on the basis of a bootstrap distribution is only approximate; one would like this probability statement to be identical to the probability statement on the basis of the sampling distribution, but for finite samples, this is not guaranteed. Moreover, there is no general statistical theory that quantifies the degree to which the bootstrap distribution approximates the sampling distribution.

It is easy to extend the bootstrap method to whole spatiotemporal data matrices, independent samples, and more than two conditions. Extending the method to whole spatiotemporal data matrices involves that Hotelling's $T^2$ statistic is replaced by the Max($T^2$) statistic. Extending the method to a study with independent samples involves that every sample (condition) has its own urn of residuals from which as many elements are drawn (with replacement) as the number of observations in this sample. The extension to a study with more than two conditions involves that, in a study with independent samples, the $T^2$ statistic is replaced by Wilk's lambda, and in a study with a within-participants design, the $T^2$ statistic is computed on a combined difference vector, as described previously.

## Applications

### *Does Diazepam Modulate the Auditory Evoked Potential?*

Jongsma, van Rijn, van Schaijk, and Coenen (2000) studied the effect of diazepam on the rat auditory evoked potential (AEP). In a within-participants experiment, Jongsma et al. obtained a single-channel ERP of eight Wistar rats in two conditions: diazepam (4.0 mg per kg, s.c.) and control. The order of the conditions (diazepam and control) was counterbalanced.

AEPs were elicited by trains of 10 repetitive tone-pip stimuli. Within a train, the interstimulus interval was 2 s. Two successive trains were separated by an interval of 4 s. Here, we only consider the AEP elicited by the first tone-pip in the train. However, essentially the same results were found for the AEPs elicited by the other tone-pips.

The EEG was bandpass filtered (between 0.1 and 500 Hz) and recorded digitally with a sampling frequency of 1,024 Hz. For every rat, the average EEG over 150 trials was calculated, separately for the experimental and the control conditions. The EEG in the interval between $-100$ and 0 ms before stimulus onset was used for baseline correction (removing the DC component). The grand averages are shown in Figure 1. From this figure, it appears that the N2 component of the AEP is much more pronounced under diazepam than in the control condition.

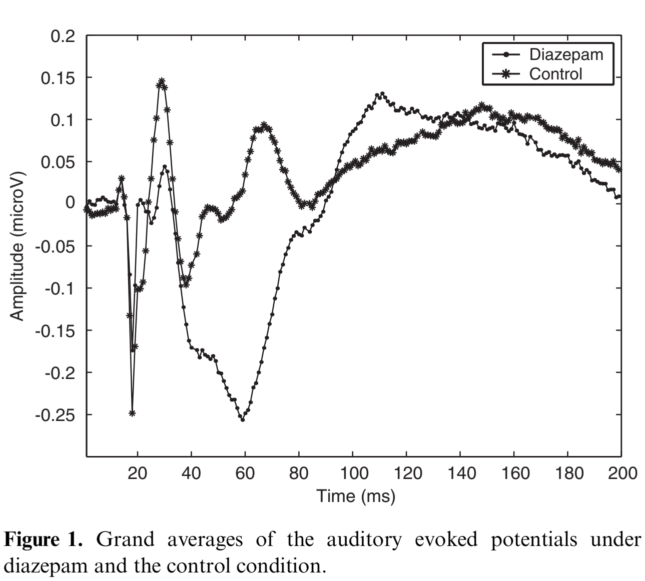

**Figure 1.** Grand averages of the auditory evoked potentials under diazepam and the control condition.

To test the statistical significance of the effect of diazepam, the Max($T^2$) randomization test was performed. Because only a single-channel potential was observed, the $T^2$ statistic is a simple squared paired samples $T$ statistic. To improve the reliability of the variance, $s^2_{\mathrm{diff}}$ in Equation 3 was replaced by a moving average (spanning a width of 31 ms) of the time-point-specific variances $s^2_{\mathrm{diff}}$. Also, because one is usually not interested in significant differences at a single time point only, a moving average (spanning a width of 10 ms) was computed on the time series of $T^2$ statistics. This resulted in a smoother time series. This time series is shown in Figure 2 as the line with the diamonds.

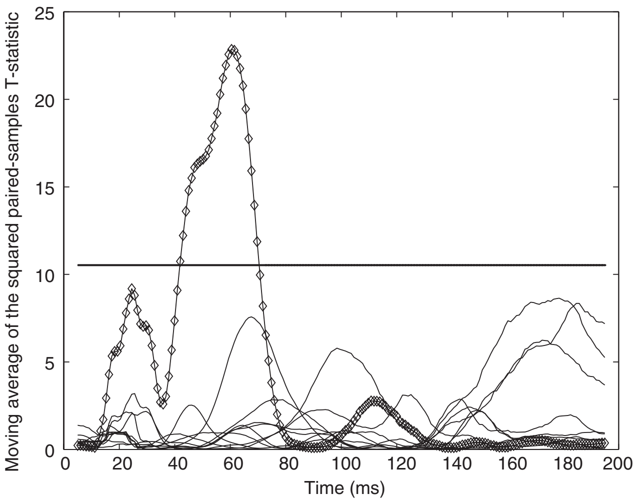

**Figure 2.** Plots of the time series of the moving average of the observed squared paired samples $T$ statistics (denoted by diamonds), the estimated 95th quantile under the randomization distribution of Max($T^2$) (denoted by the thick horizontal line), and 10 random draws of time series of squared paired samples $T$ statistics from their randomization distribution (denoted by thin lines).

The statistical test of the effect of diazepam involved the computation of the 95th quantile of the randomization distribution of Max($T^2$). This quantile was estimated by drawing 10,000 Max($T^2$) values under its randomization distribution and taking the 500th largest of these values. This estimated quantile is equal to 10.53 and it is shown in Figure 2 by the thick horizontal line. The observed time series of the squared paired samples $T$ statistics rises above the critical horizontal line after about 45 ms and remains above this line until about 72 ms poststimulus. This is in line with the fact that the estimated $p$ value of the observed Max($T^2$) value is equal to .004. Thus, it can be concluded that diazepam modulates the AEP.

In Figure 2, the thin lines are 10 random draws from the randomization distribution of the time series of the squared paired samples $T$ statistics. Of these 10 random draws, none have a maximum that is larger than the observed Max($T^2$) value.

### *The Time Course of Orthographic Processing Reflected by ERPs*

A second application involved the data of an ERP experiment that was conducted to study visual word recognition (Maris, 2002). The focus is on the comparison of the ERPs that are observed in the condition with regular words and the ERPs that are observed in the condition with pseudohomophones. Pseudohomophones are nonwords that, when pronounced, sound like an existing word (e.g., *gaim*, *werd*, *rane*). Maris was interested in the difference between the ERPs in response to regular words and those in response to pseudohomophones, because that comparison may give information on the time course of orthographic processing. More specifically, pseudohomophones differ from regular words only in that their spelling is unfamiliar (and incorrect), and therefore the time course of the difference between the ERPs may give information on the time course of orthographic processing.

The participants in the experiment by Maris (2002) performed a lexical decision (LD) task: For every stimulus, which was either an existing word or a nonword (not only pseudohomophones but also nonhomophonic pronounceable nonwords and nonpronounceable nonwords), they indicated whether it was a word or a nonword. Two versions of the LD task were used: In one version, they pressed the response button when the stimulus was a word (the go-on-word LD task) and in the other version they pressed the response button when the stimulus was a nonword (the go-on-nonword LD task). Every participant saw half of the stimuli in the go-on-word version and the other half in the go-on-nonword version of the LD task. In the following, the ERPs on the regular words in the go-on-word LD task are compared to the ERPs on the pseudohomophones in the go-on-nonword LD task. Thus, in both conditions, the same motor response was given.

The EEG was registered on 27 channels for a total of 1,200 ms: 400 ms before and 800 ms during stimulus presentation. The potentials observed in the prestimulus interval were used for baseline correction. The EEG was bandpass filtered (between 0.1 and 200 Hz) and recorded digitally with a sample frequency of 1000 Hz.

There was no hypothesis about where (i.e., on which channel) and when (i.e., in which time interval) the difference between regular words and pseudohomophones could be observed. Therefore, we used the Max($T^2$) randomization test to evaluate the statistical significance of the difference between the two 27 × 800 spatiotemporal data matrices. For the purpose of illustration, this test was applied to the data of a single participant: 18 single-trial ERPs on regular words and 21 on pseudohomophones. These trials were selected by removing all trials on which an incorrect response was given (i.e., trials on which the button was not pressed) and all trials with artifacts due to eye movements.

To improve the reliability of the estimate of the variance–covariance matrix, the $S_{\text{pooled}}$ in Equation 1 was replaced by a moving average (spanning a width of 31 ms) of the time-point-specific variance–covariance matrices $S_{\text{pooled}}$. Also, a moving average (spanning a width of 10 ms) was computed on the time series of $T^2$ statistics, which resulted in a smoother time series. This time series is shown in Figure 3.

The statistical test of the difference between the ERPs on regular words and pseudohomophones involves the computation of the 95th quantile of the randomization distribution of Max($T^2$). This quantile was estimated by drawing 10,000 Max($T^2$) values under its randomization distribution and taking the 500th largest of these values. This estimated quantile is equal to 66.832 and it is shown in Figure 3 by the thick horizontal line. The observed time series of (the moving average of) Hotelling's $T^2$ statistic rises above the critical horizontal line after about 180 ms and, in the next 75 ms, remains above this line most of the time. This is a relevant finding because, in the data of the same participant there was no early significant effect of a phonological manipulation (e.g., pronounceable vs. nonpronounceable nonwords). Thus, for at least one participant, the electrophysiological manifestion of orthographic processing is visible at an earlier stage than the electrophysiological manifestion of phonological processing.

## Conclusions

From the present study, it can be concluded that, for situations in which classical statistical tests cannot be used, the randomization test is an excellent alternative. One useful application of the randomization test is the comparison of topographies. Because the number of observations is often less than the number of channels plus 1, the classical multivariate $T$ test by means of Hotelling's $T^2$ statistic cannot be used. This problem can easily be solved by (a) replacing the variance–covariance matrix in Hotelling' $T^2$ statistic ($S_{\text{pooled}}$ for independent samples or $S_{\text{diff}}$ for

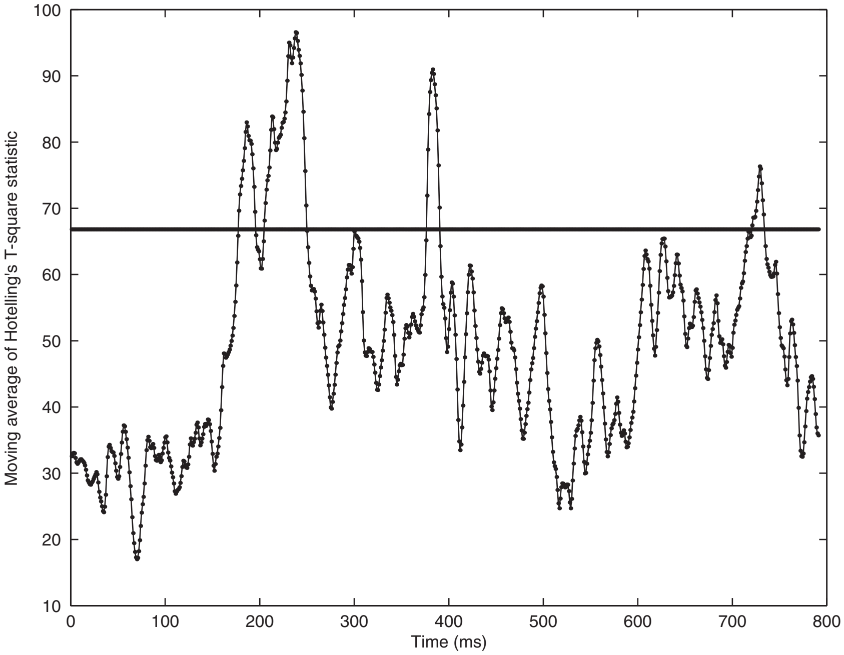

**Figure 3.** Plot of the time series of the moving average of the observed Hotelling's $T^2$ statistics, denoted by the jagged line, and the estimated 95th quantile under the randomization distribution of Max($T^2$), denoted by the horizontal line.

paired samples) by its moving average, and (b) evaluating the resulting statistic under its randomization distribution. This procedure takes advantage of the fact that an average variance–covariance matrix is more reliable than a time-specific variance–covariance matrix and therefore will result in a more sensitive statistical test (at least, if the variance–covariance matrices are constant in the time interval over which the moving average is computed).

Another useful application of the randomization test is the comparison of whole spatiotemporal data matrices. The Max($T^2$) statistic is very well suited for this comparison and the randomization distribution is a convenient reference distribution to evaluate its statistical significance.

A potential problem with the two statistical tests mentioned above is that their results may depend on the two tuning parameters: (a) the width of the moving average of the variance–covariance matrix and/or (b) the width of the moving average of the time-point-specific $T^2$ statistics. To protect against type 1 error inflation, these tuning parameters should be fixed in advance (rather than being chosen on the basis of the data). This requires scientific discipline on the part of the researcher. However, this is not essentially different from the testing of multiple comparisons in (M)ANOVA: A more conservative criterium for statistical significance is required when these comparisons are chosen on the basis of the data than when they are chosen a priori, and it is the researcher's responsibility to apply the appropriate one.

On the basis of a randomization test, one cannot generalize to a population; one can only conclude that, for this particular sample, the difference between the conditions is so large that it cannot be attributed to the randomization mechanism (which may, accidentally, assign participants with some large ERP component to one condition and the other participants to the other). I do not consider this a serious disadvantage of the randomization test because the objective of generalizing to some population of interest can only be attained if the participants are drawn at random from this population. This way of selecting particpants (specifying a population of interest and drawing at random from it) is rarely used in psychophysiological studies.

The bootstrap is another resampling method that can be used for testing the same null hypotheses that are tested by means of the randomization test. The bootstrap distribution is an approximation to the sampling distribution. In contrast to the exact probability statements that can be made on the basis of the randomization distribution, only approximate probability statements can be made on the basis of the bootstrap distribution. Moreover, there is no general statistical theory that quantifies the degree to which the bootstrap distribution approximates the sampling distribution.

## REFERENCES

Achim, A. (2001). Statistical detection of between-group differences in event-related potentials. *Clinical Neurophysiology*, *112*, 1023–1034.

Fabiani, M., Gratton, G., Corballis, P. M., Cheng, J., & Friedman, D. (1998). Bootstrap assessment of the reliability of maxima in surface maps of brain activity of individual subjects derived with electrophysiological and optical methods. *Behavior Research Methods, Instruments, & Computers*, *30*, 78–86.

Galán, L., Biscay, R., Rodríguez, J. L., Pérez-Abalo, M. C., & Rodríguez, R. (1997). Testing topographic differences between event-related brain potentials by using nonparametric combinations of permutation tests. *Electroencephalography and Clinical Neurophysiology*, *102*, 240–247.

Haig, A. R., & Gordon, E. (1995). Projection onto centroids difference vectors: A new approach to determine between group topographical differences, applied to p3 amplitude in schizophrenia. *Brain Topography*, *8*, 67–73.

Johnson, R. A., & Wichern, D. W. (1988). *Applied multivariate statistical analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Jongsma, M. L. A., van Rijn, C. M., van Schaijk, W. J., & Coenen, A. M. L. (2000). Effects of diazepam on auditory evoked potentials of rats elicited in a ten-tone paradigm. *Neuropsychobiology*, *42*, 158–162.

Karnisky, W., Blair, R. C., & Snider, A. D. (1994). An exact statistical method for comparing topographic maps with any number of subjects and electrodes. *Brain Topography*, *6*, 203–210.

Maris, E. (2002). The time-course of orthographic and phonological processing in lexical decision: An ERP-study. Manuscript in preparation.

Maxwell, S. E., & Delaney, H. D. (1990). Designing experiments and analyzing data. Pacific Grove, CA: Brooks/Cole.

Wasserman, S., & Bockenholt, U. (1989). Bootstrapping: Applications to psychophysiology. *Psychophysiology*, *26*, 208–221.

(Received March 15, 2002; Accepted June 30, 2003)