Journal of Neuroscience Methods 139 (2004) 111–120

JOURNAL OF NEUROSCIENCE METHODS

www.elsevier.com/locate/jneumeth

Multivariate tests for the evaluation of high-dimensional EEG data

Claudia Hemmelmann [a,*], Manfred Horn [a], Susanne Reiterer [b,c], Bärbel Schack [a], Thomas Süsse [a], Sabine Weiss [b,d]

[a] Institute of Medical Statistics, Computer Sciences and Documentation, University of Jena, D-07740 Jena, Germany
[b] Brain Research Institute, Cognitive Neuroscience Group, University of Vienna, Vienna, Austria
[c] Neurological Clinic, University of Tuebingen, Tuebingen, Germany
[d] SFB 360, University of Bielefeld, Bielefeld, Germany

Received 8 December 2003; received in revised form 20 April 2004; accepted 21 April 2004

Abstract

In this paper several multivariate tests are presented, in particular permutation tests, which can be used in multiple endpoint problems as for example in comparisons of high-dimensional vectors of EEG data. We have investigated the power of these tests using artificial data in simulations and real EEG data. It is obvious that no one multivariate test is uniformly most powerful. The power of the different methods depends in different ways on the correlation between the endpoints, on the number of endpoints for which differences exist and on other factors. Based on our findings, we have derived rules of thumb regarding under which configurations a particular test should be used. In order to demonstrate the properties of different multivariate tests we applied them to EEG coherence data. As an example for the paired samples case, we compared the 171-dimensional coherence vectors observed for the alpha1-band while processing either concrete or abstract nouns and obtained significant global differences for some sections of time. As an example for the unpaired samples case, we compared the coherence vectors observed for language students and non-language students who processed an English text and found a significant global difference.

Keywords: Coherence; EEG data; Global hypothesis; Language; Multiple endpoints; Multivariate test; Permutation test; Power of test

1. Introduction

The improvement of EEG/MEG measurement equipment permits the registration of a high number of channels. Additionally, modern analysis procedures yield large sets of high-dimensional parameters, which have to be evaluated statistically. Furthermore, the statistical evaluation of spectral parameters of different frequency bands becomes nearly unmanageable.

Many authors use an α-level test for each single component or endpoint of the observational vector, see e.g. Rappelsberger and Petsche (1988). However, this practice results in a large number of false positive statements. There exist several techniques to cope with this general drawback in multiple comparisons. Corresponding multiple tests will be considered in a forthcoming paper. In this paper, we will deal with so-called global tests or multivariate tests. A multivariate test provides one joint statement on all endpoints, whereas a multiple test provides a statement for each endpoint.

A well-known multivariate test is Hotelling's T^2-test. However, this test cannot be executed if the dimension of the data is higher than the number of subjects, which is the usual situation in EEG studies. Another drawback is that the T^2-test requires multivariate normal distributions. Therefore, several new statistical methodologies have been developed. Recently, several authors have proposed special test statistics in permutation tests for evaluating high-dimensional EEG data (see e.g. Karniski et al., 1994; Galan et al., 1997; Harmony et al., 2001). Permutation tests do not require special distributions and have several other advantages (see Ludbrook and Dudley, 1998).

Our paper has the following aims: (a) to give an overview of appropriate multivariate tests including novel methods, (b) to study and to compare the power characteristic of different multivariate tests using simulations with different data configurations, (c) to derive general rules for determining which test is suitable for which configuration, and (d) to

* Corresponding author. Tel.: +49-3641-9-33610; fax: +49-3641-9-33200.
E-mail address: hemmel@imsid.uni-jena.de (C. Hemmelmann).

0165-0270/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jneumeth.2004.04.013

demonstrate the use of different multivariate tests in comparisons of large sets of coherence values obtained from EEG recordings during language processing in one group of subjects under different conditions (paired samples problem) and in two groups of subjects performing the same cognitive task (two independent samples problem).

2. Methods

2.1. Multivariate tests

In analogy to the univariate case, we must differentiate between the case of paired samples and unpaired samples, i.e. two independent samples. In the paired samples case, we observe n subjects under two different conditions, say A and B. In the case of two independent samples, we observe n_1 subjects under condition A and n_2 subjects under condition B.

Our observations are vectors of dimension k. Let $\boldsymbol{x} = (x_1, \ldots, x_k)$ and $\boldsymbol{y} = (y_1, \ldots, y_k)$ denote the random vectors with means $(\mu_{x_1}, \ldots, \mu_{x_k})$ and $(\mu_{y_1}, \ldots, \mu_{y_k})$, respectively. In the paired and the unpaired samples case, $\boldsymbol{x}$ stands for the observations under condition A and $\boldsymbol{y}$ under condition B. We now can formulate the individual null hypotheses $H_1 : \mu_{x_1} = \mu_{y_1}, \ldots, H_k : \mu_{x_k} = \mu_{y_k}$ and the global hypothesis $H_0 : \mu_{x_1} = \mu_{y_1}, \ldots, \mu_{x_k} = \mu_{y_k}$. This means, H_0 is the intersection $H_0 = H_1 \cap \ldots \cap H_k$. Tests for H_0 are called global tests or multivariate tests, tests for $H_1, \ldots, H_k$ are called multiple tests.

2.1.1. The methods of Bonferroni and Simes

The simplest way to test H_0 at level α is to test all individual hypotheses $H_1, \ldots, H_k$ at level α/k and to reject H_0, if and only if, at least one H_i can be rejected. This is the well-known Bonferroni method. Let $P_1, \ldots, P_k$ be the P-values obtained when testing $H_1, \ldots, H_k$, and let $P_{(1)} \leq \ldots \leq P_{(k)}$ denote the ordered P_i. Then, the Bonferroni method rejects H_0 if $P_{(1)} \leq \alpha/k$.

Another simple test is the global test of Simes (1986). This test rejects H_0 if $P_{(i)} \leq \alpha i/k$ for at least one i ($1 \leq i \leq k$). Note, that the mathematical proof that this test keeps the α-level was given only for the case of uncorrelated P-values. However, simulations by Simes (1986), Samuel-Chan (1996) and other authors permit the conclusion that with two-sided comparisons the α-level is always kept. It is also kept with one-sided comparisons when the P-values are positively correlated. Only with negative correlations a slight but acceptable anti-conservativeness could be observed. Hence, there are no practical objections against using this test, generally.

A disadvantage of the Bonferroni method and the Simes test is that they cannot take into consideration the correlation between the endpoints or P-values. Nevertheless, with certain configurations they have higher power than some other tests that consider the correlation, see Section 3.

2.1.2. Methods of the O'Brien type

A pure multivariate test, which considers the correlation, is Hotelling's T^2-test, which is well known. However, as already mentioned in Section 1, this test is not appropriate for the analysis of EEG data, which comprise many components with few subjects. Another disadvantage is that this test requires k-variate normal distributions. The T^2-test has also the drawback that it cannot differentiate between any departures from H_0 and departures of all endpoints into the same direction, i.e. it is not very sensitive against effects in the same direction.

In order to overcome the latter disadvantage of the T^2-test, several attempts have been made, e.g. in O'Brien (1984), Pocock et al. (1987), Tang et al. (1989). The rationale behind all methods proposed is the following: For each subject, one calculates a single score from its k component values. In this way, the k-variate problem is reduced to a univariate problem. With these scores, the matched pairs t-test is executed in the paired samples situation and the two-sample t-test in the two independent samples situation. For example, the score belonging to a subject of the so-called ordinary least squares (OLS) test of O'Brien is simply the sum of its standardized component values, and OLR stands for a version that uses the corresponding ranks. Unfortunately, it turned out that with these methods the α-level could not be kept when the sample sizes are small, i.e. these tests are anti-conservative, see Kropf (2000), Reitmeir and Wassmer (1996), Frick (1997). However, Läuter et al. (1996) succeeded in constructing corresponding tests that exactly keep the α-level that are sensitive against departures from H_0 either in one direction or in both directions, and that can be used with any k and n. We will consider three different tests of Läuter et al. (1996), namely the standardized sum test (SS test), the principal component test (PC test) without scale correction and the PC test with scale correction. The SS test is sensitive against departures from H_0 in all endpoints into the same direction. With this test, the different components may be measured in different scales, e.g. one in mm, another in kg, etc. or also if the ranges of possible values are different, e.g. if the range of one endpoint is 1–10 mm and the range of another endpoint 20–50 mm. The PC test with scale correction should be used if the different endpoints are measured in different scales or if the endpoints differ in their ranges. Otherwise, the PC test without scale correction is recommended. Useful descriptions of Läuter tests and other tests of the O'Brien type can be found in Bregenzer (2000) as well as in Kropf (2000).

These tests have one of the disadvantages of the T^2-test: they require k-variate normality. Tests that do not require special distributions are permutation tests.

2.1.3. Permutation tests

The idea of permutation tests is an old one, however their broad practical application became possible only by fast computers. Below we will use numerical examples in order to explain the permutation principle for paired

Table 1
Permutations of paired samples data of dimension $k = 2$ with $n = 3$ subjects, and values of different test statistics

Permutation	Subject	(x_1, x_2)	(y_1, y_2)	t_1	t_2	t_{sum}	$t_{\|sum\|}$	t_{max}
1 (original data)	1	(8, 6)	(5, 2)	3.46	5.20	8.66	8.66	5.20
	2	(4, 3)	(3, 1)					
	3	(6, 7)	(4, 4)					
2	1	(8, 6)	(5, 2)	0.46	0.48	0.94	0.94	0.48
	2	(4, 3)	(3, 1)					
	3	(4, 4)	(6, 7)					
3	1	(8, 6)	(5, 2)	1.11	0.90	2.01	2.01	1.11
	2	(3, 1)	(4, 3)					
	3	(6, 7)	(4, 4)					
4	1	(8, 6)	(5, 2)	0.00	−0.15	−0.15	0.15	−0.15
	2	(3, 1)	(4, 3)					
	3	(4, 4)	(6, 7)					
5	1	(5, 2)	(8, 6)	0.00	0.15	0.15	0.15	0.15
	2	(4, 3)	(3, 1)					
	3	(6, 7)	(4, 4)					
6	1	(5, 2)	(8, 6)	−1.11	−0.90	−2.01	2.01	−1.11
	2	(4, 3)	(3, 1)					
	3	(4, 4)	(6, 7)					
7	1	(5, 2)	(8, 6)	−0.46	−0.48	−0.94	0.94	−0.48
	2	(3, 1)	(4, 3)					
	3	(6, 7)	(4, 4)					
8	1	(5, 2)	(8, 6)	−3.46	−5.20	−8.66	8.66	−5.20
	2	(3, 1)	(4, 3)					
	3	(4, 4)	(6, 7)					

sample and two independent sample analyses that employ t-statistics.

2.1.3.1. Paired samples. We consider an artificial numerical example given in Blair and Karniski (1993), see Table 1. The sample size is $n = 3$, the dimension of observational vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is $k = 2$.

The total number of permutations is $2^n = 2^3 = 8$. What we call permutation 1 is the original data. A permutation is obtained by exchanging the $\boldsymbol{x}$-vector of a subject for the $\boldsymbol{y}$-vector of the same subject. Note that complete $\boldsymbol{x}$- and $\boldsymbol{y}$-vectors must be exchanged as their components are correlated. In this way, permutation 2 was obtained by exchanging the $\boldsymbol{x}$-vector for the $\boldsymbol{y}$-vector of subject 3, permutation 3 by exchanging the $\boldsymbol{x}$-vector for the $\boldsymbol{y}$-vector of subject 2, permutation 4 by exchanging the $\boldsymbol{x}$-vectors for the $\boldsymbol{y}$-vectors of subjects 2 and 3, etc. This approach is justified because all such permutations are equally likely when the null hypothesis of no difference between the two conditions is true.

Now, for each permutation, the paired samples t-statistic was calculated separately for each component. t_1 is the t-value when comparing x_1 and y_1, and t_2 when comparing x_2 and y_2. In order to combine the information from the univariate test statistics in a single multivariate test statistic, we have various possibilities. We can calculate $t_{sum} = \sum_{j=1}^{k} t_j$ or $t_{|sum|} = \sum_{j=1}^{k} |t_j|$ or $t_{max} = t'_j$, where t'_j is equal to the t_j $(j = 1, \dots, k)$ which has the greatest absolute value. These test statistics were already used in Blair et al. (1994). In our example we have $t_{sum} = t_1 + t_2$, $t_{|sum|} = |t_1| + |t_2|$ and $t_{max} = t_1$ if $|t_1| > |t_2|$ or $t_{max} = t_2$ if $|t_1| \leq |t_2|$.

The different test statistics are sensitive to different specific forms of departures from H_0. t_{sum} is sensitive to departures of all endpoints in the same direction, in contrast to $t_{|sum|}$. t_{max} is sensitive to departures in only a few endpoints. Other appropriate multivariate test statistics are the t-values of the Läuter tests which are denoted by t_{SS} for the SS test and by t_{PC} for the PC test without scale correction and by t_{PC+} for the PC test with scale correction. The use of these test statistics in permutation tests was already proposed in Kropf (2000).

The statistic $t_{|sum|}$ permits only two-sided testing, in contrast to the other statistics. One-sided testing requires to specify a direction of testing in advance. In our example, one-sided testing means to show that the mean vector of x_1 and x_2 tends to be greater than that of y_1 and y_2. Two sided testing does not specify a direction.

The sampling distribution of multivariate permutation statistics is obtained by computing the desired test statistic for all permutations. As under H_0 all permutations are equally likely, we can easily calculate the P-value, i.e. the probability of obtaining an observation which under H_0 is at least as extreme as that of the original data. In our case, the P-value is the number of permutations for which the test statistic has a value not smaller than that of the original data, divided by the total number of permutations. In our example, the test statistics t_{sum}, $t_{|\text{sum}|}$, and $t_{\max}$ attain their maximum value for permutation 1, i.e. for the original data. Therefore, the probability of obtaining for t_{sum} a value not smaller than 8.66 is $1/8 = 0.125$ as only one of the eight t_{sum} values is not smaller than 8.66, i.e. the one-sided P-value is 0.125. The same applies to $t_{\max}$. For $t_{|\text{sum}|}$ we have only a two-sided test. Its P-value is $2/8 = 0.25$. In order to calculate two-sided P-value for t_{sum} (or $t_{\max}$), we determine the number of permutations where the absolute value of t_{sum} ($t_{\max}$) is equal to or greater than the corresponding absolute value for the original data, and divide it by the number of permutations. Then the two-sided P-value for t_{sum} (and $t_{\max}$) is 0.25.

If n is large, it is difficult to perform the calculations for all 2^n permutations. Then, the so-called Monte Carlo method is used which means that M permutations are randomly selected out of all possible permutations. The P-value is then the number of permutations for which the test statistic has a value not smaller than that for the original data, divided by M. The difference between the exact and the Monte Carlo method is negligible when M is large.

2.1.3.2. Two independent samples. Before we explain the permutation test for two independent samples we introduce a new multivariate test statistics for two-sided comparisons of two independent samples. Its use in permutation tests showed a surprisingly high power for some configurations, see Section 3. The suggestion comes from Good (2000) who used the term $\sum_{i=1}^{n_2}(y_i - \bar{x})^2$ in a test statistic for detecting differences between means in the univariate case for a special situation.

Let the k-variate observations we obtain for two independent samples be given in a data matrix

$$\begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n_1 1} & \cdots & x_{n_1 k} \\ y_{11} & \cdots & y_{1k} \\ \vdots & & \vdots \\ y_{n_2 1} & \cdots & y_{n_2 k} \end{pmatrix}.$$

For all components, we calculate the sample means $\bar{x}_j = (1/n_1)\sum_{i=1}^{n_1} x_{ij}$ and $\bar{y}_j = (1/n_2)\sum_{i=1}^{n_2} y_{ij}$, the sample variances $s_{x_j}^2 = (1/(n_1-1))\sum_{i=1}^{n_1}(x_{ij} - \bar{x}_j)^2$ and $s_{y_j}^2 = (1/(n_2-1))\sum_{i=1}^{n_2}(y_{ij} - \bar{y}_j)^2$, and the differences $v_{ij} = (y_{ij} - \bar{x}_j)/\sqrt{s_{y_j}^2 + s_{x_j}^2/n_1}$ and $w_{ij} = (x_{ij} - \bar{y}_j)/\sqrt{s_{x_j}^2 + s_{y_j}^2/n_2}$ ($j = 1, \ldots, k$). Note that the denominators in v_{ij} and w_{ij} are the estimates of the standard deviations of the respective numerators. Now we calculate the means $\bar{v} = (1/kn_2)\sum_{j=1}^{k}\sum_{i=1}^{n_2} v_{ij}$ and $\bar{w} = (1/kn_1)\sum_{j=1}^{k}\sum_{i=1}^{n_1} w_{ij}$, the differences $a_{ij} = v_{ij} - \bar{v}$ ($i = 1, \ldots, n_2$, $j = 1, \ldots, k$) and $a_{ij}^* = w_{ij} - \bar{w}$ ($i = 1, \ldots, n_1$, $j = 1, \ldots, k$), and the component-wise test statistics $a_j = (1/n_2)\sum_{i=1}^{n_2} a_{ij}^2 + (1/n_1)\sum_{i=1}^{n_1}(a_{ij}^*)^2$ ($j = 1, \ldots, k$). Finally we determine $t_a = \max\{a_1, \ldots, a_k\}$, which is the new global test statistic we propose. Note that a_j ($j = 1, \ldots, k$) and with it t_a are symmetric in x and y, i.e. exchanging the two samples gives the same value.

The rationale behind this proposal will be explained under simplifying assumptions. Let the mean vectors $(\mu_{x_1}, \ldots, \mu_{x_k})$ and $(\mu_{y_1}, \ldots, \mu_{y_k})$ differ in exactly m components, for example $\mu_{y_i} - \mu_{x_i} = \Delta_i$ ($i = 1, \ldots, m$) and $\mu_{y_i} - \mu_{x_i} = 0$ ($i = m+1, \ldots, k$). Assume that n_1 and n_2 are very large, so that $s_{x_j}^2$ and $s_{y_j}^2$ can be replaced by their respective expectations $\sigma_{x_j}^2$ and $\sigma_{y_j}^2$, and $s_{x_j}^2/n_1$ and $s_{y_j}^2/n_2$ by 0. In addition assume that $\sigma_{x_j}^2 = \sigma_{y_j}^2 = \sigma_j^2$. Then

$$E(a_{ij}) = \frac{E(y_{ij}) - E(\bar{x}_j)}{\sigma_j} - E(\bar{v}) = \frac{\Delta_j}{\sigma_j} - \frac{1}{k}\sum_{l=1}^{m}\frac{\Delta_l}{\sigma_l},$$

and

$$E(a_{ij}^*) = \frac{E(x_{ij}) - E(\bar{y}_j)}{\sigma_j} - E(\bar{w}) = -\left(\frac{\Delta_j}{\sigma_j} - \frac{1}{k}\sum_{l=1}^{m}\frac{\Delta_l}{\sigma_l}\right) \quad (j = 1, \ldots, m).$$

Thus, for large values of $|\Delta_j/\sigma_j|$ we can expect that $|a_{ij}|$ and $|a_{ij}^*|$ with it a_j and finally t_a are large.

Note that $E(a_{ij}) = E(a_{ij}^*) = 0$, if $m = k$ and $\Delta_1/\sigma_1 = \cdots = \Delta_k/\sigma$, so that t_a provides a low power. This can be seen in our simulation results in Figs. 3 and 4 in Section 3. However, such a situation will rarely be met in practice.

We now explain the permutation principle when comparing two independent samples. Again, we use a simple numerical example, see Table 2. The two independent samples have the sizes $n_1 = 3$ (subjects under condition A) and $n_2 = 2$ (subjects under condition B). The dimension of the observational vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is $k = 2$. Permutation 1 is the original data. The way of permuting is different from that in the paired samples case. We now exchange vectors observed in subjects under A for vectors observed in (other) subjects under B. So permutation 2 was obtained by exchanging the $\boldsymbol{x}$-vector of the first subject of the sample under A for the $\boldsymbol{y}$-vector of the first subject of the sample under B, permutation 3 by exchanging the $\boldsymbol{x}$-vector of the second subject under A for the $\boldsymbol{y}$-vector of the first subject under B, etc. In this way we obtain $(n_1 + n_2)!/(n_1!n_2!) = 5!/(3!2!) = 10$ permutations. The justification for this approach is similar as in the paired samples case. Now, for each permutation, the values t_1 and t_2 of the t-statistic for the case of two independent samples were calculated separately for each of the two components. Finally, the multivariate test statistics t_{sum}, $t_{|\text{sum}|}$, and $t_{\max}$ were calculated from t_1 and t_2 in the same way as in the paired samples case. In addition, the new test statistic t_a was calculated. Now, the P-values can be calculated as described for the paired samples case. For

Table 2
Permutations of two-dimensional data of two independent samples with sizes $n_1 = 3$ and $n_2 = 2$, and values of different test statistics

Permutation	(x_1, x_2)	(y_1, y_2)	t_1	t_2	t_{sum}	$t_{\|sum\|}$	t_{max}	t_a
1 (original data)	(8, 6) (4, 3) (6, 7)	(5, 2) (3, 1)	1.20	2.40	3.60	3.60	2.40	1.72
2	(5, 2) (4, 3) (6, 7)	(8, 6) (3, 1)	−0.25	0.18	−0.07	0.43	−0.25	0.82
3	(8, 6) (5, 2) (6, 7)	(4, 3) (3, 1)	2.37	1.42	3.79	3.79	2.37	1.22
4	(8, 6) (4, 3) (5, 2)	(6, 7) (3, 1)	0.61	−0.12	0.59	0.73	0.61	0.97
5	(3, 1) (4, 3) (6, 7)	(5, 2) (8, 6)	−1.36	−0.12	−1.48	1.48	−1.36	1.22
6	(8, 6) (3, 1) (6, 7)	(5, 2) (4, 3)	0.61	0.89	1.50	1.50	0.89	0.78
7	(8, 6) (4, 3) (3, 1)	(5, 2) (6, 7)	−0.25	−0.44	−0.69	0.69	−0.44	0.77
8	(5, 2) (3, 1) (6, 7)	(8, 6) (4, 3)	−0.71	−0.44	−1.15	1.15	−0.71	0.84
9	(5, 2) (4, 3) (3, 1)	(8, 6) (6, 7)	−2.85	−5.40	−8.25	8.25	−5.40	3.94
10	(8, 6) (5, 2) (3, 1)	(4, 3) (6, 7)	0.17	−0.81	−0.64	0.98	−0.81	1.10

example, as we have two permutations with $t_{sum} \geq 3.60$, the one-sided P-value for t_{sum} is $2/10 = 0.2$. The one-sided P-value for t_{max} is 0.1. The two-sided P-values for t_{max} and t_a are 0.2, and for t_{sum} and $t_{|sum|}$ are 0.3.

2.2. Artificial data and real data (EEG data)

2.2.1. Artificial data

We generated samples from k-variate normal, exponential and log-normal distributions for special configurations of means and correlation coefficients, and executed the different multivariate tests described in Section 2.1. The components of k-variate normally distributed vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ had common variance 1. For all distributions, the means $\mu_{x_1}, \ldots, \mu_{x_k}$ and $\mu_{y_1}, \ldots, \mu_{y_k}$ were chosen so, that $\mu_{y_i} - \mu_{x_i} = \Delta$ $(i = 1, \ldots, m)$ and $\mu_{y_i} - \mu_{x_i} = 0$ $(i = m+1, \ldots, k)$. This means we considered m false and $k-m$ true hypotheses, and the deviations of the false hypotheses were all into the same direction. The value of m was varied between 1 and k.

In the paired samples case, we denote the coefficients of correlation between the differences $x_i - y_i$ and $x_j - y_j$ by ρ_{ij} $(1 \leq i \leq j \leq k)$. In the case of two independent samples, ρ_{ij} denotes the coefficient of correlation between the components x_i and x_j as well as between the components y_i and y_j $(1 \leq i \leq j \leq k)$. For both paired and unpaired samples, we first considered the case $\rho_{ij} = \rho$ $(i \neq j)$, i.e. constant (low, moderate or high) correlation. However, in most practical situations, the correlation coefficients ρ_{ij} do not have the same value and the same sign. Therefore, we also considered the following two types of correlation matrices Corr1 and Corr2. The matrix

$$\mathrm{Corr1} = \begin{pmatrix} 1 & \frac{k-1}{k} & \frac{k-2}{k} & \cdots & \frac{1}{k} \\ \frac{k-1}{k} & 1 & \frac{k-1}{k} & \cdots & \frac{2}{k} \\ \frac{k-2}{k} & \frac{k-1}{k} & 1 & \cdots & \frac{3}{k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \frac{1}{k} & \frac{2}{k} & \frac{3}{k} & \cdots & 1 \end{pmatrix}$$

may be typical for longitudinal observations, e.g. time series where neighboring observations have higher correlations than more distant observations. The matrix

$$\mathrm{Corr2} = \begin{pmatrix} R1 & R2 & R2 \\ R2 & R1 & R2 \\ R2 & R2 & R1 \end{pmatrix}$$

with

$$R1 = \begin{pmatrix} 1 & 2/3 & \cdots & 2/3 \\ 2/3 & 1 & \cdots & 2/3 \\ \vdots & \vdots & \ldots & \vdots \\ 2/3 & 2/3 & \cdots & 1 \end{pmatrix}$$

and

$$R2 = \begin{pmatrix} -1/3 & -1/3 & \cdots & -1/3 \\ -1/3 & -1/3 & \cdots & -1/3 \\ \vdots & \vdots & \ldots & \vdots \\ -1/3 & -1/3 & \cdots & -1/3 \end{pmatrix}$$

was used in order to investigate a case where both, positive and negative correlations occur.

The number of repeated simulations for any configuration was 5000, the number of permutations in each permutation test was 1000.

2.2.2. EEG data

The EEG was recorded with 19 gold-disc electrodes according to the 10–20 system against the averaged signals $(A1+A2)/2$ of both ear lobe electrodes. Filter settings were 0.3–35 Hz, sampling frequency was 256 Hz. Epochs with artifacts were eliminated from further processing.

2.2.2.1. Word processing under different conditions. Twenty-three right-handed female native German speakers participated in the experiment (age 22–26). Two word lists each containing either 25 concrete or 25 abstract German nouns were selected and psycholinguistically controlled (e.g. Weiss and Müller, 2003). Nouns had a mean word length of $0.9\,\mathrm{s} \pm 0.08$ and were auditorily presented with an interstimulus interval of 2.5 s. Participants were requested to memorize the presented nouns and had to recall them after the presentation. A trigger marked the beginning of each word presentation, and the following 1 s EEG epochs were selected for coherence analysis. Instantaneous coherence values for each electrode pair of the alpha1-band (8–10 Hz) were computed with a time resolution of approximately 4 ms in accordance with the sampling frequency of 256 Hz on the basis of an ARMA model with time-varying model parameters (see e.g. Schack et al., 1999, 2003). The chosen frequency resolution was 0.5 Hz. Averaging was performed within the alpha1-band (8–10 Hz) over adjacent spectral lines. For each trial 250 instantaneous coherence values were obtained. The next step was averaging over the single trials time point by time point. For data reduction, the 1 s trials were divided into 10 sections of 100 ms. This yielded 10 coherence values per presentation mode and per word category.

The comparison of coherence values for all 171 electrode pairs accompanying concrete and abstract nouns processing requires a multivariate test for paired samples.

2.2.2.2. Foreign language processing by different groups. A group of 19 foreign language students (all studying English language and linguistics at university level) and a second group of 19 non-language students (studying other subjects than languages), all mother-tongue German speakers, participated in the experiment. The subjects had to watch various video sequences of TV news on a TV screen. The stimulus material used in this study comprised nine individual tasks. Each task was a sequence of an approximately 2 min video-recorded TV news text. The TV news texts were presented in British English, in American English and in Austrian German, and in three modalities: visual + auditory (TV mode), auditory only and visual only. Looking at a gray flickering picture was chosen as a control situation in order to control individual base-line effects.

For all electrode pairs ordinary coherence values of the alpha1-frequency band (8–10 Hz) were calculated according to the Bartlett algorithm by using 2 s intervals for averaging procedures.

We evaluated one of the three TV-news (British English) which were presented in TV mode (visually + auditory). The comparison of the 171-dimensional vectors of coherence values of the two groups of subjects requires a multivariate test for two independent samples.

3. Results

3.1. Power characteristics and power comparisons

Power is a statistical term. The power of a multivariate test is the probability of rejecting the global null hypothesis if it is false. It may depend on various quantities, e.g. the correlation between the components, the number of components, the number of components where differences exist, etc. In order to estimate the power and to study its dependence on these quantities it is common to perform computer simulations. For this purpose, we generated samples from k-variate distributions as described in Section 2.2.1, and applied the different multivariate tests introduced in Section 2.1. The results reported in the sequel are restricted to two-sided testing and to normal distributions. The results for other distributions were similar. They are not shown here.

We observed that the power of most (but not all) methods increases with increasing m/k. This is what one would expect.

When $\rho_{ij} = \rho$ for $i \neq j$, i.e. when the correlation is the same for all pairs of components, the influence of the correlation on the power is different for different methods, see Figs. 1 and 2.

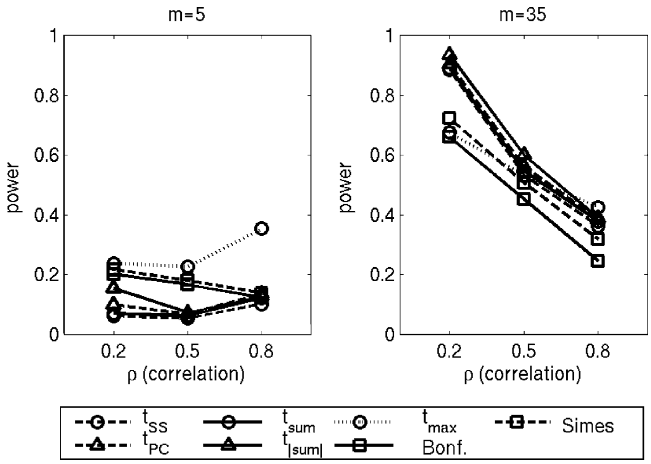

Fig. 1. Power dependence on the correlation ρ in the paired samples case for $m = 5$ (left side) and $m = 35$ (right side) ($k = 40$, $n = 20$, $\Delta = 0.4$, $\alpha = 0.05$).

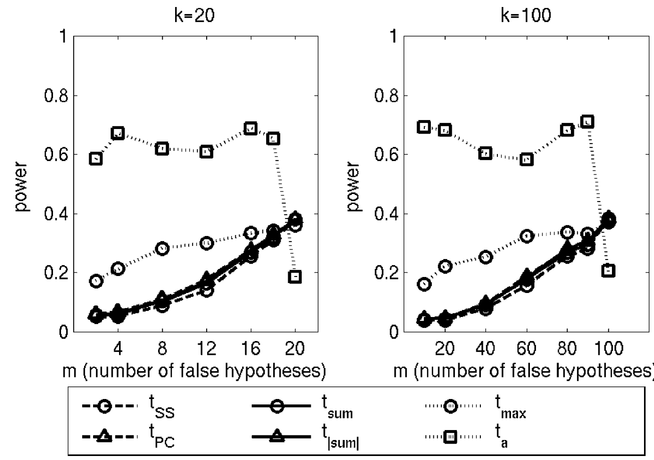

Fig. 3. Power values in the two independent samples case for $k = 20$ (left side) and $k = 100$ (right side) ($\rho = 0.8$, $n_1 = n_2 = 10$, $\Delta = 0.7$, $\alpha = 0.05$).

The power of the permutation test with $t_{\max}$ is higher than the power of all other methods, if m/k is small, see Fig. 1, left side. A comparison of the remaining methods shows that the Simes method and the Bonferroni method have a relatively high power if ρ is small or medium. A different characteristic and rank order of the methods can be observed if m/k is large, see Fig. 1, right side. Note also that with large m/k the power decreases when ρ increases. This tendency was also reported in other papers, e.g. in Blair et al. (1994).

$t_{\max}$ is superior also with medium m/k as long as the correlation is high, see Fig. 2, right side. However, $t_{\max}$ and also Bonferroni and Simes are inferior with low correlation and large m/k, see Fig. 2, left side. Note that in each situation a better test exists than the methods of Simes or Bonferroni. This applies also for the two independent samples case. Hence, in the power plots for this case we will omit the Simes and Bonferroni methods.

In the two independent samples case, we obtained similar results as in the paired samples case when using the same test statistics. When applying the new statistic t_a introduced in Section 2.1.3.2, we obtained more powerful tests for equal correlations, see Fig. 3. Note that the power of the test with t_a is higher for $k = 100$ than for $k = 20$ when m/k is 0.1 or 0.9, whereas the power of the other tests seems to be independent of k. Only, when $m = k$, t_a provides a low power, see also Fig. 4. The explanation is given in Section 2.1.3.2.

When considering the correlation structures given by the matrices Corr1 and Corr2, we obtained the power estimates in Fig. 4. Under Corr1 the power curves do not differ so strongly as under Corr2 where the tests with t_{SS} and t_{sum} are distinctly better. Under both correlation structures t_a provides low power. However, as already mentioned, in these simulations all m mean differences $\mu_{y_i} - \mu_{x_i}$ were positive. If we considered positive and negative differences, the power for t_{SS} and t_{sum} would be lower and the power for t_a higher.

Table 3 provides an overview of which methods were best in which configuration. The statements there may be helpful

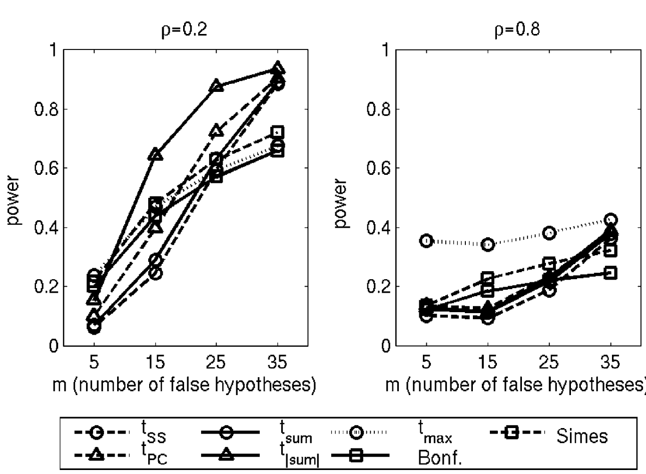

Fig. 2. Power dependence on m/k in the paired samples case for $\rho = 0.2$ (left side) and $\rho = 0.8$ (right side) ($k = 40$, $n = 20$, $\Delta = 0.4$, $\alpha = 0.05$).

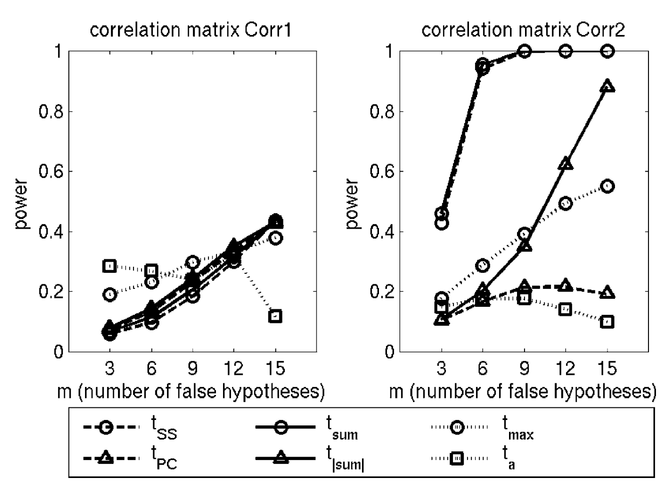

Fig. 4. Power values in the two independent samples case with correlation matrix Corr1 (left side) and correlation matrix Corr2 (right side) ($k = 15$, $n_1 = n_2 = 10$, $\Delta = 0.7$, $\alpha = 0.05$).

Table 3
Best multivariate tests under different configurations for paired samples and two independent samples

Correlation	Relative number of false hypotheses	Paired samples	Two independent samples
$\rho = 0.2$	Small	$t_{\max}$, Simes, Bonferroni	t_a
	Medium	$t_{\lvert\text{sum}\rvert}$	$t_{\lvert\text{sum}\rvert}$
	Large	$t_{\lvert\text{sum}\rvert}$, t_{sum}, t_{SS}, t_{PC}	$t_{\lvert\text{sum}\rvert}$, OLS, t_{SS}
$\rho = 0.8$	Small	$t_{\max}$	t_a
	Medium	$t_{\max}$	t_a
	Large	$t_{\max}$, $t_{\lvert\text{sum}\rvert}$, t_{sum}	t_{SS}, $t_{\lvert\text{sum}\rvert}$
Corr1	Small	$t_{\max}$	t_a
	Medium	$t_{\max}$	$t_{\max}$
	Large	t_{sum}, $t_{\lvert\text{sum}\rvert}$, t_{SS}, t_{PC}	t_{sum}, $t_{\lvert\text{sum}\rvert}$, t_{SS}, t_{PC}
Corr2	Small	t_{sum}, t_{SS}	t_{sum}, OLS, t_{SS}, OLR
	Medium	t_{sum}, t_{SS}	t_{sum}, OLS, t_{SS}, OLR
	Large	t_{sum}, t_{SS}, $t_{\lvert\text{sum}\rvert}$	t_{sum}, OLS, t_{SS}, OLR

to decide which of the different tests or test statistics should be used for evaluating data at hand.

3.2. Applications of multivariate tests to EEG coherence data

3.2.1. Paired samples of word processing under two different conditions

The data we now evaluate come from the experiment described in Section 2.2.2.1. The number of subjects was 23. For each subject, a vector of 171 coherence values was obtained under two different conditions, namely under the processing of either concrete or abstract nouns. The coherence values are given for 10 sections of time: 0–100, 100–200, ..., 900–1000 ms. For each section, we tested the global hypothesis of no difference between concrete and abstract nouns processing in all 171 components. The *P*-values obtained for the permutation test with the different test statistics are given in Table 4. *P*-values smaller than 0.05 are in bold type, which means that the corresponding tests rejected the global hypothesis at the 5% level. The *P*-values for the Bonferroni and the Simes methods are omitted. Neither rejected the global hypothesis. The pure Läuter tests also did not provide significance in contrast to permutation tests that used the Läuter test statistics. It can be seen that in different sections different tests provided significance.

3.2.2. Paired samples from longitudinal observations of word processing for only one component under two different conditions

In this section, we consider the same data as in the former section. However, we now evaluate only the coherence values obtained for one of the 171 components, namely for the electrode pair P3/O1, as a function of time. As we have 10 sections of time, we have a multivariate test problem of dimension $k = 10$. The mean coherence values under processing of abstract and concrete nouns are given for the 10 sections of time in Fig. 5, left side.

This illustration and also the boxplots of the pairwise differences in Fig. 5, right side, suggest that the presentation of abstract nouns results in higher coherence values than the presentation of concrete nouns for this electrode pair. When we separately test each of the 10 differences with the Wilcoxon test, we obtain *P*-values greater than 0.05, see Fig. 5, right side. However, the *P*-values of the multivariate permutation tests with the test statistics t_{sum} and t_{SS} are 0.05 and 0.046, respectively, see Table 5. This permits the rejection of the global null hypothesis and confirms that the curve of coherence values over the 10 sections of time

Table 4
P-values of the multivariate permutation test with different test statistics for 10 sections of time

Test	Time in ms									
	0–100	100–200	200–300	300–400	400–500	500–600	600–700	700–800	800–900	900–1000
t_{sum}	**0.0164**	0.1202	0.1216	0.3455	**0.0344**	**0.0482**	0.3091	0.3433	0.6591	0.5607
$t_{\lvert\text{sum}\rvert}$	**0.0212**	0.1462	0.0842	0.1910	**0.0230**	**0.0308**	0.2198	0.2743	0.4269	0.5005
$t_{\max}$	0.2845	**0.0368**	0.0660	0.0576	**0.0252**	**0.0174**	**0.0458**	**0.0336**	0.2517	0.7197
t_{SS}	**0.0144**	0.1218	0.1290	0.3575	**0.0336**	0.0526	0.3133	0.3477	0.6591	0.5581
t_{PC}	**0.0138**	0.0928	0.0640	0.1616	**0.0202**	**0.0208**	0.2250	0.2887	0.5581	0.5167
$t_{\text{PC+}}$	**0.0148**	0.1022	0.0732	0.2216	**0.0236**	**0.0308**	0.2835	0.3285	0.5617	0.5143

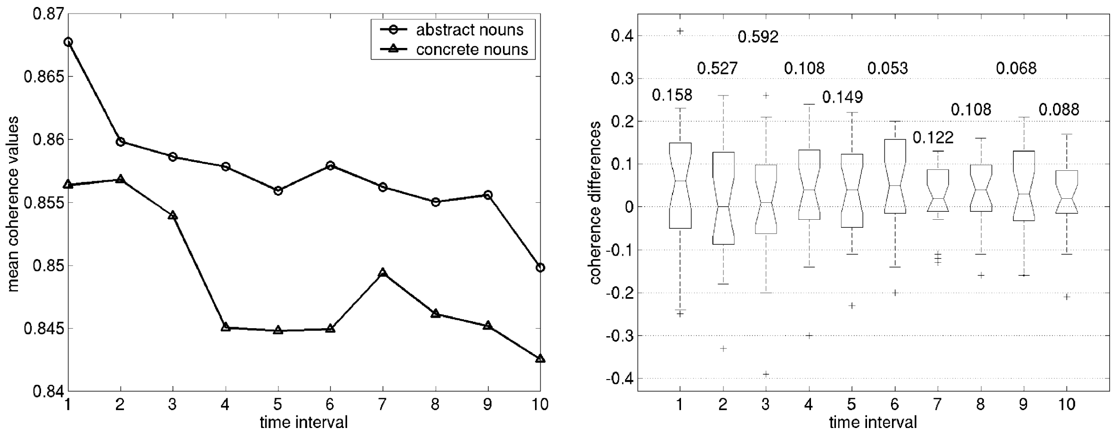

Fig. 5. Mean coherence values for the electrodes P3 and O1 for abstract and concrete nouns processing (left side) and corresponding boxplots and Wilcoxon test P-values of coherence differences (right side).

Table 5
P-values of the multivariate permutation test with different test statistics when comparing coherence values of the electrode pair P3/O1 over 10 sections of time

Test statistics	P-value
t_{sum}	**0.050**
$t_{\lvert\text{sum}\rvert}$	0.062
t_{max}	0.300
t_{SS}	**0.046**
t_{PC}	0.137
$t_{\text{PC+}}$	0.058

is significantly higher for abstract nouns than for concrete nouns.

3.2.3. Independent samples data of foreign language processing by two different groups

In this section, we evaluate data of the experiment as described in Section 2.2.2.2. The aim is to compare coherence values of the language students and the non-language students. Again, for each student a vector of 171 coherence values is given. We have to test the global null hypothesis that the vectors of both groups have the same expectations. The P-values of the permutation test with different test statistics are given in Table 6. At the 5% level, we can reject the global null hypothesis of no difference between the two groups with the test statistics t_{sum}, $t_{\lvert\text{sum}\rvert}$, t_{max}, t_{SS}, and t_a.

Table 6
P-values for the multivariate permutation test with different test statistics when comparing coherence values of language and non-language students

Test statistics	P-value
t_{sum}	**0.037**
$t_{\lvert\text{sum}\rvert}$	**0.049**
t_{max}	**0.026**
t_{SS}	**0.035**
t_{PC}	0.072
$t_{\text{PC+}}$	0.052
t_a	**0.048**

4. Discussion

With multiple endpoints, two different aims are of interest: (1) to test globally whether there is an overall effect when considering all endpoints simultaneously and (2) to provide an efficacy statement for each individual endpoint. We need so-called multivariate or global tests for the first aim and multiple tests for the second aim. In this paper we only deal with multivariate tests. It should be mentioned that multiple tests frequently fail to detect effects for individual endpoints, whereas a multivariate test provides significance. This means that different endpoints may jointly explain a treatment effect when using a multivariate test.

In this paper, we have introduced several multivariate tests with an emphasis on multivariate permutation tests. The computational effort of permutation tests is relatively high. But permutation tests have various advantages. Good (2000) states: ‘Permutation tests permit us to choose the test statistic best suited to the task at hand’.

All simulation results presented in this paper were obtained under k-variate normality. With exponential and log-normal distributions we obtained similar results as with normal distributions. Hence, they are not reported in this paper.

Our investigations of the varying power characteristics showed that among the different multivariate tests or test statistics, no specific one can be determined as the best. It depends, on the correlation structure, on the number of endpoints for which differences exist, on the magnitude of the differences, on the total number of endpoints and on other factors, which test is able to reject the global hypothesis. Thus, prior knowledge concerning the correlation between the components and the relative number of components at which differences exist should be used to find an appropriate multivariate test. Advice and recommendations as to which methods are appropriate can be found in Table 3 of Section 3.1.

In order to demonstrate the properties of different multivariate tests we applied them to EEG coherence data obtained while participants processed either different word categories or an English text.

Firstly, multivariate tests for paired samples were applied to the processing of concrete and abstract nouns. It is well known that processing of concrete and abstract nouns differs in various ways, which has been assessed with several different neuro-physiological and -psychological techniques (e.g. Bleasdale, 1987; Coltheart, 1987; Kiehl et al., 1999; Weiss et al., 1999). With several of our tests, we obtained significant global differences between the 171-dimensional coherence vectors for concrete and abstract nouns for the time intervals between 0–200 and 400–800 ms. These findings partly correspond to previous results on phase relations which showed the most prominent differences between concrete and abstract nouns processing between 300–500 and 600–800 ms, respectively (Schack et al., 2003). The time intervals showing the most significant differences coincide with the results of an ERP-study of West and Holcomb (2000), who noticed that concrete words elicited a more negative ERP than abstract words between 300 and 800 ms.

Secondly, multivariate tests for independent samples were applied to EEG coherence data from high and low proficiency second language speakers processing English TV news reports. It is a commonsense argument that people differ in their proficiency and amount of training in their second languages, hence have different levels of performance. EEG coherence has been used to detect the differences between high and low foreign language performers (Reiterer and Rappelsberger, 2001). A significant global difference between the 171-dimensional vectors of coherence values (intensity of coherence) of language students and non-language students could be found with most of our multivariate tests. Obviously, this was caused by differences for most electrode pairs with a tendency of larger coherence values for language students.

Acknowledgements

We would like to thank the two referees for their useful comments and suggestions. This work was supported by Interdisziplinäres Zentrum für Klinische Forschung Jena, Projekt 1.8, by the Austrian Science Foundation (“Herta Firnberg”—project T127) and the German Science Foundation (SFB 360).

References

Blair RC, Karniski W. An alternative method for significance testing of waveform difference potentials. Psychophysiology 1993;30:518–24.

Blair RC, Higgins JJ, Karniski W, Kromrey JD. A study of multivariate permutation tests which may replace Hotelling’s T^2 test in prescribed circumstances. Multivariate Behav Res 1994;29:141–63.

Bleasdale FA. Concreteness-dependent associative priming: separate lexical organization for concrete and abstract words. J Exp Psychol: Learn Mem Cogn 1987;13:582–94.

Bregenzer T. Tests zur Auswertung klinischer Studien mit multiplen Endpunkten. Doctoral Thesis, University Cologne; 2000.

Coltheart M. Deep dyslexia: a review of the syndrome. In: Coltheart M, Patterson KE, Marshall JC, editors. Deep dyslexia. London: Routledge and Kegan Paul; 1987.

Frick H. A note on the bias of O’Brien’s OLS test. Biometrical J 1997;39:125–8.

Galan L, Biscay R, Rodriguez JL, Perez-Abalo MC, Rodriguez R. Testing topographic differences between event related brain potentials by using non-parametric combinations of permutation tests. Electroencephalogr Clin Neurophysiol 1997;102:240–7.

Good P. Permutation tests. New York: Springer; 2000.

Harmony T, Fernandez T, Fernandez-Bouzas A, Silva-Pereyra J, Bosch J, Diaz-Comas L, et al. EEG changes during word and figure categorization. Clin Neurophysiol 2001;112:1486–98.

Karniski W, Blair RC, Snider AD. An exact statistical method for comparing topographic maps. Brain Topogr 1994;6:203–10.

Kiehl KA, Liddle PF, Smith AM, Mendreck A, Forster BB, Hare RD. Neural pathways involved in the processing of concrete and abstract words. Hum Brain Mapp 1999;7:225–33.

Kropf S. Hochdimensionale multivariate Verfahren in der medizinischen Statistik. Aachen: Shaker Verlag; 2000.

Läuter J, Glimm E, Kropf S. New multivariate tests for data with an inherent structure. Biometrical J 1996;38:5–23.

Ludbrook J, Dudley H. Why permutation tests are superior to t and F tests in biomedical research. Am Statistician 1998;52:127–32.

O’Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics 1984;40:1079–87.

Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. Biometrics 1987;43:487–98.

Rappelsberger P, Petsche H. Probability mapping: power and coherence analyses of cognitive processes. Brain Topogr 1988;1:46–54.

Reiterer S, Rappelsberger P. EEG-coherence analysis and foreign language processing. Neuroimage 2001;13:592.

Reitmeir P, Wassmer G. One-sided multiple endpoint testing in two-sample comparisons. Commun Stat Simulation Comput 1996;25:99–117.

Samuel-Chan E. Is the Simes improved Bonferroni procedure conservative? Biometrika 1996;83:928–33.

Schack B, Rappelsberger P, Weiss S, Möller E. Adaptive phase estimation and its application in EEG analysis of word processing. J Neurosci Methods 1999;93:49–59.

Schack B, Weiss S, Rappelsberger P. Cerebral information transfer during word processing: where and when does it occur and how fast is it? Hum Brain Mapp 2003;19:18–36.

Simes RJ. An improved Bonferroni procedure for multiple tests of significance. Biometrika 1986;73:751–4.

Tang DI, Gnecco C, Geller NL. An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. Biometrika 1989;76:577–83.

Weiss S, Müller HM. The contribution of EEG coherence to the investigation of language. Brain Lang 2003;85:325–43.

Weiss S, Müller HM, Rappelsberger P. Processing concepts and scenarios: electrophysiological findings on language representation. In: Riegler A, Peschl M, Stein A, editors. Understanding representation in the cognitive sciences. New York: Plenum Press; 1999. p. 237–45.

West WC, Holcomb PJ. Imaginal, semantic, and surface-level processing of concrete and abstract words: an electrophysiological investigation. J Cogn Neurosci 2000;12:1024–37.