Psychophysiology, ■■■ (2011), 1–12. Wiley Periodicals, Inc. Printed in the USA.
Copyright © 2011 Society for Psychophysiological Research
DOI: 10.1111/j.1469-8986.2011.01272.x

Mass univariate analysis of event-related brain potentials/fields II: Simulation studies

DAVID M. GROPPE,[a] THOMAS P. URBACH,[a] AND MARTA KUTAS[a,b]

[a]Department of Cognitive Science, University of California, San Diego, La Jolla, California, USA
[b]Department of Neurosciences, University of California, San Diego, La Jolla, California, USA

Abstract

Mass univariate analysis is a relatively new approach for the study of ERPs/ERFs. It consists of many statistical tests and one of several powerful corrections for multiple comparisons. Multiple comparison corrections differ in their power and permissiveness. Moreover, some methods are not guaranteed to work or may be overly sensitive to uninteresting deviations from the null hypothesis. Here we report the results of simulations assessing the accuracy, permissiveness, and power of six popular multiple comparison corrections (permutation-based control of the familywise error rate [FWER], weak control of FWER via cluster-based permutation tests, permutation-based control of the generalized FWER, and three false discovery rate control procedures) using realistic ERP data. In addition, we look at the sensitivity of permutation tests to differences in population variance. These results will help researchers apply and interpret these procedures.

Descriptors: EEG/ERP, MEG, False discovery rate, Permutation test, Hypothesis testing

As detailed in a companion article (Groppe, Urbach, & Kutas, 2011), a mass univariate analysis is an approach to analyzing data involving a massive number of univariate hypothesis tests (e.g., *t* tests) with relatively powerful corrections for the large number of comparisons like permutation tests or false discovery rate control. This approach is superior to conventional mean time window, ANOVA-based analysis of event-related brain potentials (ERPs) and event-related brain magnetic fields (ERFs) in that it requires fewer *a priori* assumptions and can provide greater temporal and spatial resolution. While these benefits come with some loss of statistical power, in many cases, the power that remains may be more than adequate to detect effects of interest. Thus, mass univariate analyses can be a valuable complement to and, in some cases, may even obviate the need for conventional ERP/ERF analyses.

At the same time, however, some popular mass univariate procedures may not be generally appropriate for ERP/ERF analysis given that the procedures may be more permissive than desired due to the structure of ERP/ERF data or random variation in the performance of the procedure. Specifically, there are the following potential concerns:

1. Independent samples permutation tests (e.g., Blair & Karniski, 1993; Korn, Troendle, McShane, & Simon, 2004; Maris & Oostenveld, 2007) may be overly sensitive to differences in variance between the two populations being sampled. For example, the data from a clinical population (e.g., children diagnosed with attention-deficit hyperactivity disorder) may be noisier than those from the control population but otherwise identical. Although such a difference is likely not of interest, it could lead to significant test results that are misattributed to differences in central tendency (i.e., mean ERP/ERF amplitude) that are of primary interest.
2. The Benjamini and Hochberg (1995) false discovery rate (FDR) control procedure may not accurately control FDR since ERP/ERF data at one time point and sensor may be negatively correlated with ERP/ERF data at another time point and sensor. ERP/ERF data are probably generally approximately normally distributed, and the Benjamini and Hochberg FDR control procedure is not guaranteed to work on normally distributed data with such negative correlations (Benjamini & Yekutieli, 2001). However, since FDR of normally distributed data tends to behave as if the tests are independent as the number of tests increases (Clarke & Hall, 2009), the Benjamini and Hochberg procedure may typically be accurate in practice.
3. The Benjamini, Krieger, and Yekutieli (2006) FDR control procedure may not accurately control FDR since ERP/ERF data at one time point and sensor are typically highly

The authors would like to thank Ryan Canolty, Sandy Clarke, Ed Korn, and Bradley Voytek for help with researching this article. This research was supported by U.S. National Institute of Child Health and Human Development grant HD22614 and National Institute of Aging grant AG08313 to Marta Kutas and a University of California, San Diego Faculty Fellows fellowship to David Groppe. All data used for the research reported in this manuscript were collected from human volunteers who were 18 years of age or older and participated in the experiments for class credit or pay after providing informed consent. The University of California, San Diego Institutional Review Board approved the experimental protocol.

Address correspondence to: David M. Groppe, Department of Cognitive Science, 0515, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093-0515. E-mail: dgroppe@cogsci.ucsd.edu

PSYCHOPHYSIOLOGY

correlated, and this procedure is only guaranteed to work when the different dimensions of the data are independent. Again, though, in practice this procedure may typically be accurate (Clarke & Hall, 2009).
4. FDR control procedures may be prone to alarmingly high proportions of false discoveries, even if on average they accurately control FDR. Korn and colleagues (2004), for instance, found that when the Benjamini and Hochberg procedure was applied to simulated data using an FDR level of 10%, there was a 10% chance that the true proportion of false discoveries was 29% or more, for some simulation parameters.
5. Since the cluster-based permutation test (Bullmore et al., 1999; Maris & Oostenveld, 2007) provides only weak control of the familywise error rate (FWER), it is more likely, albeit to an unknown degree, to make false discoveries than methods that provide strong control of FWER when some effects are truly present. Thus, the FWER and FDR of the procedure may be surprisingly high when effects are truly present.

In this paper, we report the results of simulation studies designed to evaluate the extent to which these potential problems are actual problems for mass univariate analyses of ERPs/ERFs. To that end, we applied all of the multiple comparison correction procedures reviewed in the companion article (Groppe et al., 2011) to realistic, simulated ERP data, with which we could precisely evaluate their susceptibility to the concerns raised above as well as the relative power of these methods to detect various sized effects. Specifically the procedures we investigate here are:

- Permutation-based strong control of FWER based on the standard t statistic ($t_{\max}$—Blair & Karniski, 1993),
- Cluster-based permutation tests with weak control of FWER based on maximum cluster-level mass (Bullmore et al., 1999; Maris & Oostenveld, 2007),
- Benjamini and Hochberg's (1995) FDR control algorithm (BH),
- Benjamini and Yekutieli's (2001) FDR control algorithm (BY),
- Benjamini, Kreiger, & Yekutieli's (2006) FDR control algorithm (BKY),
- Korn et al.'s (2004) permutation-based procedure for generalized familywise error rate (GFWER) control based on the standard t statistic (KTMS).

We also contrasted these methods to the Bonferroni-Holm procedure[1] (BonH—Holm, 1979) that provides strong control of FWER in order to provide a sense of how all these procedures compare to a classic multiple comparison correction with which many researchers are familiar.

Simulation Studies of the Effect of Between-Population Differences in Variance on Permutation Tests

Background

The purpose of the simulations described in this section was to assess how sensitive the independent samples permutation test is to between-population differences in variance using realistic electroencephalogram (EEG) background noise and a realistic number of comparisons. Previous work using simulated normally distributed data has shown that the independent samples permutation test based on the standard independent samples t statistic is rather insensitive to differences in variance between populations when the sizes of the samples from the two populations are equal (Murphy, 1967). This is true even when the group standard deviations differ by as much as a factor of four. However, when group sizes differ by a factor of two, the permutation test can be quite anticonservative (if the smaller sample has greater variance) or overly conservative (if the larger sample has greater variance)[2]. Note, however, that Murphy's simulations were all based on a single dependent variable (i.e., there was only a single comparison in the family of tests), and it is not clear how well these results generalize to the large number of highly correlated comparisons that would be typical of ERP/ERF analyses.

To find out if this is the case, we extended Murphy's simulations using realistic, simulated ERP data and a realistic number of comparisons. As with Murphy's original study, different numbers of participants per sample and multiple degrees of between-population variation differences were modeled. In addition, we utilized permutation tests based on three different test statistics to determine if some statistics were more insensitive than others to between-population differences in variance. These statistics were the standard independent samples t statistics (t), Welch's approximate t statistic (t_W), and the difference between group t scores[3] (t_{dif}). The equations for these statistics are as follows:

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{1}$$

$$t_W = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} \tag{2}$$

$$t_{dif} = t_1 - t_2 = \frac{\hat{\mu}_1}{\hat{\sigma}_1} - \frac{\hat{\mu}_2}{\hat{\sigma}_2} \tag{3}$$

where $\hat{\mu}_1$and $\hat{\mu}_2$ are the estimated means, $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are the estimated standard deviations, and n_1 and n_2 are the sample sizes of Sample 1 and Sample 2, respectively. $\hat{\sigma}_p$ is the estimated pooled standard deviation:

$$\hat{\sigma}_p = \sqrt{\frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}} \tag{4}$$

Based on Murphy's univariate results, we expected Welch's t to be less sensitive to differences in variance than the standard t statistic. We also expected the t_{dif} statistic to be less sensitive to differences in variance than the standard t statistic since each group of data is independently normalized by its estimated standard deviation before the between-population difference is measured.

1. Bonferroni-Holm is a slight variation of the standard Bonferroni procedure that is applicable whenever the Bonferroni procedure is appropriate and is always at least as, if not more, powerful.

2. It is worth noting that Murphy also found that the standard independent samples t test was as badly affected as the permutation test by between-population differences in variance.

3. t scores are derived by dividing the estimated mean of a population by the estimated standard deviation. Often such statistics are called z scores, but they are more accurately called t scores since the values will follow a t distribution if the data come from a normally distributed population or the sample size is sufficiently large.

Simulation Parameters

Realistic EEG background noise was derived from the data of 37 volunteers who performed a simple tone counting task (Groppe et al., 2010). The University of California, San Diego Institutional Review Board approved the experimental protocol. Each participant's EEG was recorded at 26 scalp channels using a left mastoid reference and an analog bandpass filter of 0.016–100 Hz. EEG was digitized at a 250 Hz sampling rate. After recording, the EEG was rereferenced to the algebraic mean of both mastoids, low-pass filtered at 50 Hz, and artifact polluted trials were either rejected or artifact corrected using independent components analysis (Lee, Girolami, & Sejnowski, 1999). ERPs were derived from epochs of EEG time-locked to tones and lasting from 100 to 920 ms peritone onset. The ERP was then subtracted from each epoch to produce trials of zero mean EEG background noise. On average, there were 66 trials per participant ($SD = 6$). Data points of interest for these simulations were all time points from 100 to 900 ms at all scalp channels. The median standard deviation of the background noise at these data points across all 37 participants was 9.54 μV (IQR = 3.82 μV). For further experiment details, see Groppe et al. (2010).

To simulate a single ERP experiment, participants were randomly selected (without replacement) from the pool of 37 participants and assigned to one of the two experimental groups. The ERPs were derived from each participant by randomly selecting (with replacement) 49 of that participant's background noise trials, removing the mean prestimulus voltage (− 100 to 0 ms), and averaging the trials. For each experiment, a permutation test was applied to detect differences at all scalp channels from 100 to 900 ms (201 time points × 26 channels = 5226 comparisons) using a familywise alpha level of 5%. Five thousand random permutations were used for each test to approximate the set of all possible permutations. Simulated experiments were run using different group sizes and different noise levels. To manipulate the noise level of a group, that group's ERPs were multiplied by one of the following factors: 4/3, 5/3, or 2. For each group size and noise level, 4000 simulated experiments were run to estimate the probability of erroneously rejecting one or more null hypotheses.

In addition to the above simulations, an additional set of 2000 simulations were run on a simulated N1 ERP effect (Naatanen & Winkler, 1999) to assess the power of the three different test statistics. Simulated ERPs were produced by the same procedure used above, but now a 3 μV difference between groups was added from 100 to 140 ms at 14 fronto-central electrodes. Thus, 154 of the 5226 null hypotheses were false. The size of the effect was chosen such that all tests would have a medium degree of effect sensitivity. For these simulations, there were no between-group differences in variance but the numbers of participants per group were varied as before.

Simulation Results

Figure 1 shows the FWER of the three test statistics for different sample sizes as a function of their differences in standard deviation. When using the standard t statistic and Welsh's t, the simulations replicate Murphy's three general findings:

1. When the two samples are of equal size, there is a moderate increase in FWER when the populations differ in variance.
2. When the two samples differ in size and the population corresponding to the smaller sample varies more, the permutation test tends to be anticonservative. The degree of FWER inflation can be severe (e.g., greater than 25% of the nominal level).
3. When the two samples differ in size and the population corresponding to the smaller sample varies less, the permutation test tends to be overly conservative.

The degree of FWER inaccuracy, though, is somewhat less when the permutation test is based on Welch's t rather than the standard t and the samples differ in size. The permutation test based on the t_{dif} statistic is also subject to some FWER inaccuracy when the groups differ in variance. In general, it becomes increasingly anticonservative as the difference in variance between populations increases (regardless of which sample is smaller). However, the degree of anticonservativeness is quite small, especially relative to the other two statistics.

With regards to power, the N1 effect simulation results (Figure 2) show that the standard t statistic is most powerful, with Welch's t being only somewhat less powerful. The t_{dif} statistic was considerably less powerful than the other two statistics. These differences in power tend to grow as the difference between sample sizes increases.

Simulation Studies Evaluating the Permissiveness and Power of Strong FWER, Cluster-Based Weak FWER, GFWER, and FDR Control Methods

To explore the relative permissiveness and power of the four types of multiple comparison correction methods covered in this paper, exemplars of each method were applied to four types of simulated ERP effects: a null effect (i.e., just EEG background noise), a very focal early sensory effect, a broadly distributed late effect, and the combination of the early and late effects. The motivation for simulating all four effects was to assess the frequency of false and true discoveries of the various methods as a function of the number of false null hypotheses. In addition, the simulations were performed using bimastoid referenced ERP background noise as well as average reference ERP background noise because the bimastoid reference is representative of a great proportion of ERP research and the average reference induces a large proportion of negative correlations between tests (Luck, 2005). The latter is of interest to determine if the BH and BKY methods can accurately control FDR despite a realistically large proportion of negative correlations.

This work builds on some previous evaluations of the power and permissiveness of these mass univariate procedures using simulated and real EEG/MEG data.

Previous Work on FDR Permissiveness

Regarding the accuracy of two of Benjamini and colleagues' FDR procedures and their propensity to return a high proportion of false discoveries, Hemmelmann, Horn, Susse, Vollandt, and Weiss (2005) applied the BH and BKY procedures to multivariate normal data with various covariance structures and numbers of false null hypotheses. They found that both algorithms reliably controlled FDR at or below the nominal 5% rate, even when the dimensions of the data were highly positively correlated or exhibited a mixture of positive and negative correlations. This

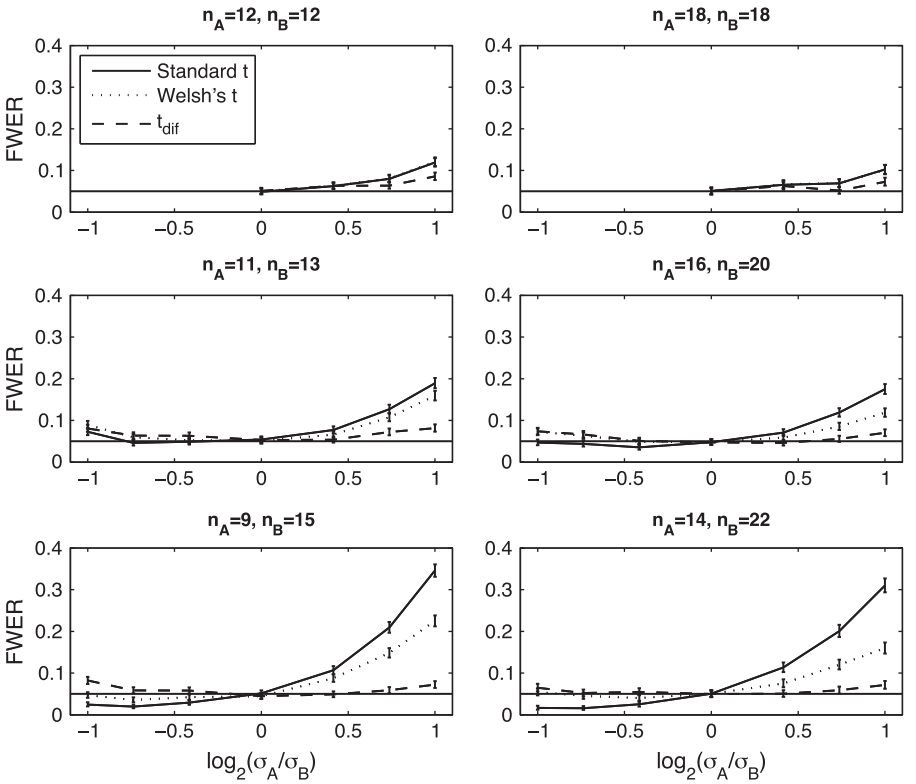

Figure 1. Familywise error rate of independent samples t_{max} permutation tests based on different t statistics as a function of between-population differences in standard deviation. n_A and n_B equal the number of subjects in Samples A and B, respectively. σ_A and σ_B equal the standard deviation of the data in Samples A and B, respectively. Left and right columns show results from simulations based on 24 and 36 total subjects, respectively. Error bars indicate 95% confidence intervals. Horizontal solid line indicates nominal 5% familywise error rate. Note that the standard t and Welsh's t statistic are mathematically equivalent when the two groups being compared consist of equal numbers of subjects.

occurred even though the number of comparisons was relatively small, 40. This result is consistent with applications of the BKY procedure to simulated data with positive correlations (Benjamini, Krieger, & Yekutieli, 2006) and of the BKY and BH procedures to simulated data with positive and negative correlations (Kim & van de Wiel, 2008). Hemmelmann and colleagues also found that the BH and BKY procedures were not terribly prone to a high proportion of false positives. Specifically, using a nominal FDR level of 5%, they found that the probability of getting 10% or more false discoveries was generally less than 20%.

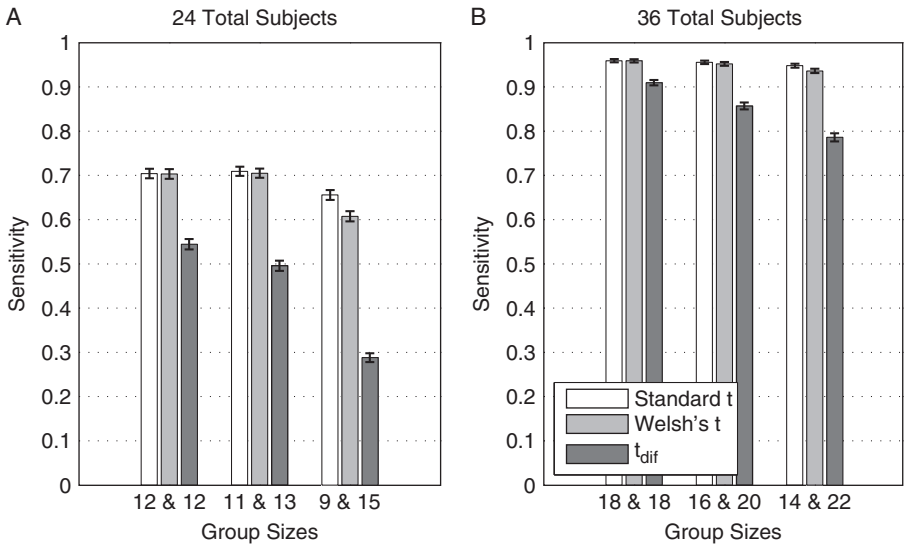

Figure 2. Sensitivity of between-group t_{max} procedure based on three different test statistics to a simulated N1 effect. A sensitivity of 1 means that the entirety of the N1 effect was detected (i.e., at all time points and electrodes). A sensitivity of 0 means that the N1 effect was entirely missed. x-axis indicates the number of subjects in each of the two samples. Error bars indicate 95% confidence intervals. Note that the standard t and Welsh's t statistic are mathematically equivalent when the two groups being compared consist of equal numbers of subjects.

While these results generally support the use of the BH and BKY procedures, it is difficult to tell how representative Hemmelmann and colleagues' simulations are of ERP/ERF data. For example, ERP/ERF data may exhibit negative correlations that are much stronger and more frequent than that of their simulated data, and this may deleteriously affect FDR control permissiveness. Lage-Castellanos, Martinez-Montes, Hernandez-Cabrera, and Galan (2010) took steps to remedy this shortcoming by applying the BH and BY algorithms to simulated ERP data using realistic ERP noise generated from actual ERP data and sine wave-derived ERP effects. Using these data, relatively small-to-moderate numbers of comparisons (200 to 3,800), and a nominal FDR level of 5%, they found that the BY algorithm always tended to control the FDR level well below 5% and that the BH algorithm generally tended to do so save for one simulation in which the estimated FDR level was too high, 9.7%. However, these results are hard to interpret since there was not a clear distinction between time points and electrodes that did and did not exhibit an ERP effect. Specifically, Lage-Castellanos et al. defined time points exhibiting ERP effects as those for which the amplitude of the ERP effect was greater than or equal to 0.5 standard deviations of the ERP noise. Since the sine-wave ERP effects differed from zero at many more time points than this, it is possible that the BH algorithm still accurately controlled FDR. In addition, Lage-Castellanos and colleagues did not estimate the propensity of these procedures to return a high proportion of false discoveries.

The work presented in this article builds on these results by using realistic simulated ERP data with a clear distinction between data points with and without ERP effects. We also evaluate all three of Benjamini and colleagues' algorithms for their ability to accurately control FDR and for their propensity to return a high proportion of false discoveries.

Previous Work on Cluster-Based Permutation Test Permissiveness

To the best of our knowledge, this is the first work to examine the FWER and FDR of cluster-based permutation tests when effects are truly present.

Previous Work on the Relative Power of Various Mass Univariate Procedures

To the best of our knowledge, there are only two published comparisons of the relative power of various mass univariate tests using ERP/ERF data. Maris and Oostenveld (2007) compared the ability of various test statistics to detect the N400m (the MEG analog of the N400 ERP effect) in real data from a single subject using a very large number of comparisons (151 sensors $\times$ 600 time points = 90,600 comparisons) and permutation tests. They explored four mass univariate statistics: the maximum sum of t scores in a cluster, the maximum size of a cluster, the maximum t score (i.e., t_{max}), and the maximum absolute difference between means.[4] The latter two, noncluster-based statistics provide strong control of FWER, and the cluster-based statistics provide weak control of FWER. Of these, they found that the two cluster-based statistics tended to be able to detect at least part of the N400m effect with the fewest number of trials, 11. The t_{max} test required approximately twice as many trials, and the maximum absolute difference between the means required about three times as many trials. It is difficult to know how reliable these differences are since Maris and Oostenveld reported quantitative results for only a single random selection of trials. However, they mention that the results were highly similar when they repeated the comparison with different sets of random trials, and there is good *a priori* reason to expect that cluster-based statistics are better at detecting a broadly distributed effect like the N400m (Bullmore et al., 1999).

More recently, Lage-Castellanos and colleagues (2010) compared the power of two of Benjamini and colleagues' FDR control algorithms (BH and BY) and the t_{max} permutation procedure on simulated ERP data and a real P3 dataset.[5] The number of comparisons ranged from 200 to 4940. On the simulated data, they found that the BH procedure tended to be 10% to 45% more powerful than t_{max}, while the BY procedure was 5% to 15% more powerful than t_{max}. When applied to real ERP data, the BH procedure again proved clearly most powerful, though the t_{max} procedure was actually more powerful than the BY procedure.

Lage-Castellanos et al.'s results are generally consistent with Hemmelmann and colleagues' (2005) analysis of various multiple comparison correction procedures using simulated normally distributed data and a real EEG coherence dataset (171 comparisons). Of the methods examined in our investigations, Hemmelmann et al. tested the BH and BKY FDR control procedures, and "step-up" variants[6] of t_{max} and Korn et al.'s GFWER procedure, which are more powerful and computationally intensive than the t_{max} and GFWER procedures used here. Across a variety of different types of covariance matrices, numbers of comparisons, and proportion of false null hypotheses, they found that FDR control was more powerful than the t_{max} and GFWER control procedures, except when a very small proportion of null hypotheses were false. They also found that the BKY procedure was generally as or more powerful than BH, especially when a large proportion of null hypotheses were false, and that their GFWER procedure was generally significantly more powerful than the step-up version of t_{max}. The relative power of these methods was qualitatively similar when applied to real EEG coherence data.

4. In addition, Maris and Oostenveld investigated four other "global statistics": sum of the positive/negative t scores, mean of the positive/negative t scores, sum of the positive/negative difference between means, and mean of the positive/negative difference between means. These statistics are "global" in that they can only detect that a difference exists somewhere in a family of tests, but they do not determine exactly which tests are significant. In contrast, the mass univariate tests identify precisely which tests are significant in the family of tests. It is worth noting that Maris and Oostenveld found that all of the mass univariate tests tended to be more powerful than the global tests. Hemmelmann et al. (2004) also found that mass univariate permutation tests were quite powerful relative to such global permutation tests when applied to artificial normally distributed data and EEG coherence data.

5. Lage-Castellanos and colleagues also investigated the relative power of local-FDR control (Efron, 2004), another method for multiple comparison correction not covered here.

6. The step-up variants of the strong FWER and GFWER procedures work analogously to the Bonferroni-Holm procedure. The permutation test is initially run including all comparisons. If no comparisons are significant, the procedure stops. Otherwise, the most significant comparison is assigned the p value from the test and removed from further analysis. This procedure (starting with another permutation test) is then repeated on the remainder of the comparisons, et cetera, until none of the comparisons still under consideration reach significance or until all comparisons are deemed significant.

Here we build on this work by evaluating all of these multiple comparison correction methods using realistic simulated ERP data with clearly defined broadly and/or narrowly distributed effects. By using realistic simulations, our results are more likely to generalize to actual ERP/ERF analyses than those of Hemmelmann et al. (2005). Moreover, we contrast some multiple comparison correction procedures that have not yet been contrasted on ERP/ERF data (e.g., cluster-based tests versus FDR controls). Also, by using a variety of simulated effects, we can better evaluate the general power of the cluster-based tests than Maris and Oostenveld (2007).

Based on previous findings and the mechanics of these methods, we expect FDR control and the cluster-based permutation test to be the most powerful procedure when broad effects are in the data and t_{max} and GFWER control to be the most powerful when only narrow effects are present. What is not clear is how the cluster-based test compares to the ability of FDR control to capture broad effects, nor just how large the differences in power are between the different methods when applied to ERP/ERF data. Understanding the latter is obviously key to deciding if using a more powerful method (e.g., BH) is worth the added uncertainty that might accompany that procedure (e.g., uncertainty as to the significance of any single test result).

Simulation Parameters

Realistic EEG background noise was derived from the data of 23 volunteers who performed a linguistic priming task (Groppe, Choi, Topkins, & Kutas, 2009). EEG recording and artifact correction parameters were the same as those used in the previous simulation study and, again, the University of California, San Diego Institutional Review Board approved the experimental protocol.

ERPs were derived from epochs of EEG time-locked to text primes and lasting from −100 to 920 ms peritone onset. The ERP was then subtracted from each epoch to produce trials of zero mean EEG background noise. On average, there were 223 trials per participant ($SD = 12$). Data points of interest for these simulations were all time points from 100 to 900 ms at all 26 scalp channels for a total of 5226 dependent variables (i.e., 26 channels × 201 time points). The median standard deviation of the background noise at these data points across all 23 participants was 10.56 μV (IQR = 3.41 μV).

To create average reference noise, the procedure was the same save that after artifact correction the mean voltage across all channels was removed from the EEG at each time point. The median standard deviation of the average reference background noise at the data points of interest across all 23 participants was 6.49 μV (IQR = 2.33 μV).

To simulate a single ERP experiment, ERPs were derived for each of the 23 participants by randomly selecting (with replacement) 49 of that participant's background noise trials, removing the mean prestimulus voltage (−100 to 0 ms), and averaging the trials. Save for simulations of ERP null effects, a deflection was added to each participant's noise ERPs to simulate one of three possible ERP effects. The first of these effects was a focal early effect based on the N170 ERP component to text (Bentin, Mouchetant-Rostaing, Giard, Echallier, & Pernier, 1999). This simulated "N170" effect was a deflection of 1 μV at a single left lateral occipital electrode from 140 to 190 ms, which is probably as extremely focal an ERP/ERF effect as one might observe. Thus for the N170 simulations, the null hypothesis was false at only 13 dependent variables (i.e., 0.2% of the total number of null hypotheses). The magnitude of the N170 effect was chosen such that the sensitivity of the multiple comparison correction procedures was medium to low.

The second simulated effect was a broad, late effect roughly based on the P3 ERP component to text (Bentin et al., 1999). The simulated "P3" effect was a deflection of 1.3 μV at 13 central and posterior electrodes from 400 to 700 ms. Thus for the P3 simulations, the null hypothesis was false at 988 dependent variables (i.e., 18.9% of the total number of null hypotheses). The magnitude of the P3 effect was chosen to be greater than that of the N170 effect but small enough to produce only a medium-to-high degree of effect sensitivity. The combined early and late effect simulations simply combined the N170 and P3 deflections. Thus for the combined effect simulations, the null hypothesis was false at 1001 dependent variables (i.e., 19.2% of the total number of null hypotheses).

Each of the four types of ERP effects were simulated 1000 times and analyzed each time by applying two-tailed, one-sample t tests to all scalp channels from 100 to 900 ms. The following seven methods were used for multiple comparison correction:

1.–2. To strongly control the FWER at 5%, the Bonferroni-Holm (Holm, 1979) and t_{max} permutation (Blair & Karniski, 1993) procedures were used.
3. To weakly control FWER at 5%, the maximum cluster-level mass permutation procedure was used (Bullmore et al., 1999; Maris & Oostenveld, 2007).
4. To control the GFWER with a 5% chance of more than one false discovery, the permutation-based procedure introduced by Korn et al. (2004) was used.
5.–7. To control FDR at 5%, the BH, BY, and BKY procedures were used. Note that because the BH and BY procedures should actually control FDR at 5% times the proportion of null hypotheses that are actually true (Benjamini & Yekutieli, 2001), their nominal level of FDR control is approximately 4% when applied to the P3 and combined N170/P3 effects due to the moderate fraction of false null hypotheses.

Finally, the degree of correlation between variables was estimated by computing the correlation between each pair of dependent variables (i.e., 13,652,925 pairs) for each of the 1000 null effect simulations. For the bimastoid referenced noise, the median percentage of negatively correlated variable pairs per simulation was 13.1% (IQR = 8.2%). For the average reference noise, the median percentage of negatively correlated variable pairs per simulation was 50.9% (IQR = 0.1%).

Simulation Results

Figures 3–6 illustrate various measures of the permissiveness of the multiple comparison correction procedures. In general, the t_{max}, cluster-based test, and KTMS procedures allow the most false discoveries when the proportion of false null hypotheses is small or none (i.e., for the N170 and null effects). However, when there are a moderate number of false null hypotheses (i.e., for the P3 and combined P3/N170 effects), the FDR control and cluster-based methods are the most permissive, with the BH and BKY methods being the most permissive of all. Importantly, all correction procedures accurately control the specified number or

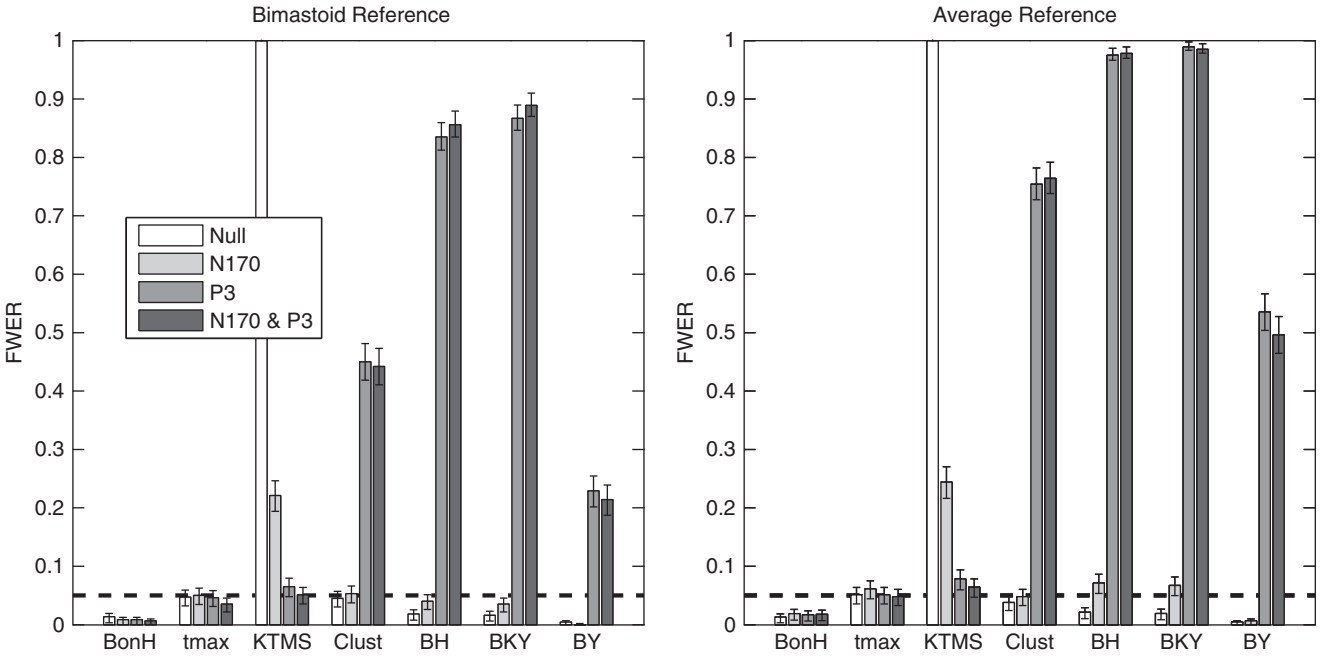

Figure 3. Familywise error rate of seven different methods of multiple comparison correction applied to four different simulated ERP effects. Dashed line indicates nominal level (5%) of FWER control for BonH and t_{max} methods for all effects. 5% is also the nominal FWER for the cluster-based test and all FDR controls when no effect is present. Error bars indicate 95% confidence intervals.

proportion of false discoveries at or below their nominal levels. This is particularly notable for the BH and BKY FDR control procedures since they are not guaranteed to control FDR due to the correlations between some pairs of variables. It is also notable that the FDR and cluster procedures only rarely produce a large percentage of false discoveries (i.e., greater than 20%—Figure 6), even though the BH procedure has been shown to do the opposite on less realistic data (Korn et al., 2004). Finally, with regard to the cluster-based permutation test, its permissiveness is comparable to that of t_{max} when few or no null hypotheses are false. When broad effects are present in the data, the cluster procedure's propensity to produce some false discoveries is

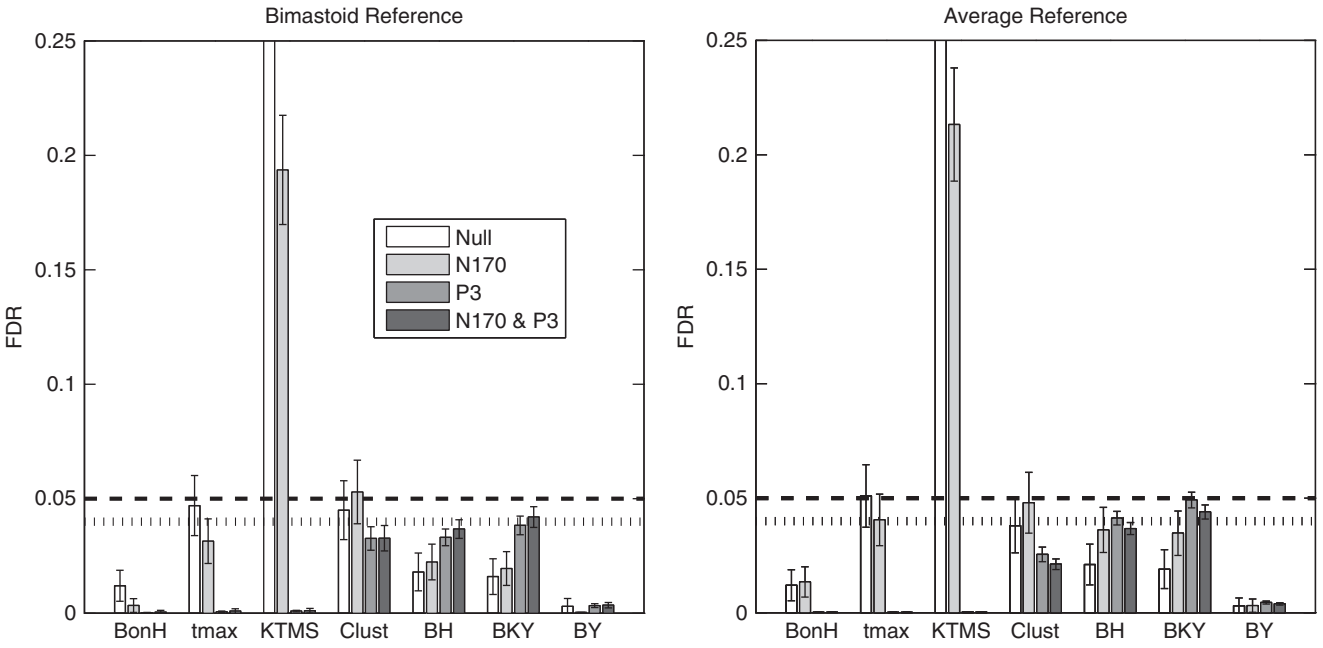

Figure 4. False discovery rate of seven different methods of multiple comparison correction applied to four different simulated ERP effects. Nominal level of FDR control for method BKY is 5% (indicated by line composed of horizontal dashes). Nominal level of FDR control for BH and BY methods is approximately 5% for Null and N170 effects and approximately 4% (indicated by line composed of vertical dashes) for P3 and combined N170/P3 effects. Error bars indicate 95% confidence intervals.

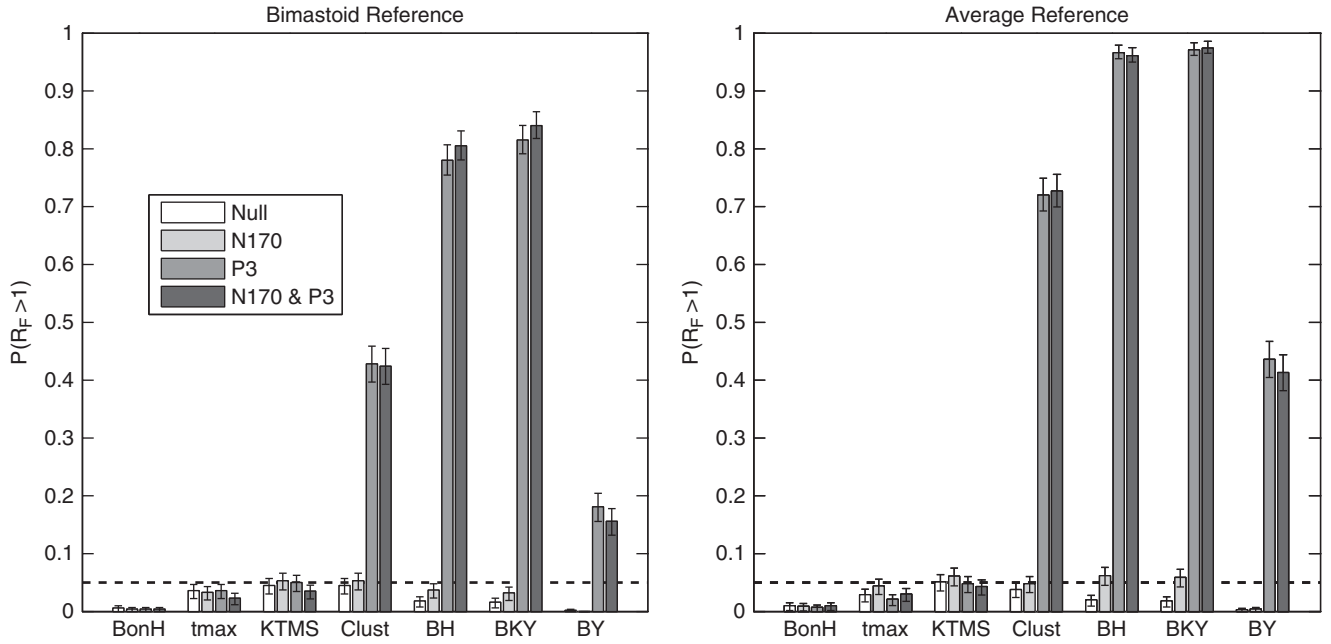

Figure 5. Probability of more than one false discovery in the family of tests for seven different methods of multiple comparison correction applied to four different simulated ERP effects. Dashed line indicates nominal level (5%) of false discovery control for KTMS method. Error bars indicate 95% confidence intervals.

considerably less than that of the BH and BKY FDR control procedures (Figures 3 & 5), and its FDR rate is comparable to that of BH and BKY (Figure 4).

Figures 7 and 8 illustrate two different measures of test power. For the narrowly distributed N170 effect, the cluster-based test has almost no ability to detect the effect, and the BY procedure is worse at detecting it than Bonferroni-Holm. The remaining procedures are about twice as likely to detect the effect than Bonferroni-Holm, with the KTMS and $t_{\max}$ methods tending to be the best. When the broad P3 effect is in the data, the cluster-based test and the FDR control methods are clearly the best at detecting the greatest proportion of tests where some effect is present, with the cluster-based procedure being the very best and the BH and BKY methods being almost as good (Figure 7). These results are generally true of each method's ability to detect at least one member of each effect as well, except for the cluster-based test, which almost never detects both the N170 and P3 because it is so poor at detecting the N170 (Figure 8).

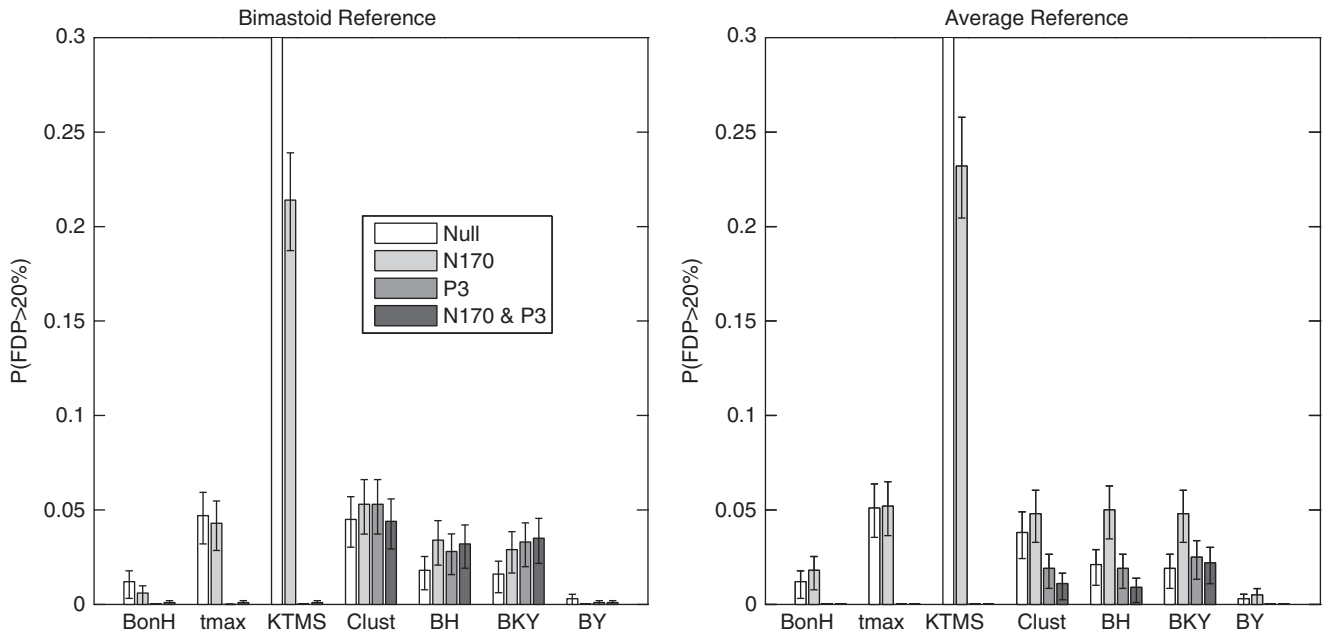

Figure 6. Probability that the proportion of discoveries that are false discoveries exceeds 20% for seven different methods of multiple comparison correction applied to four different simulated ERP effects. Error bars indicate 95% confidence intervals.

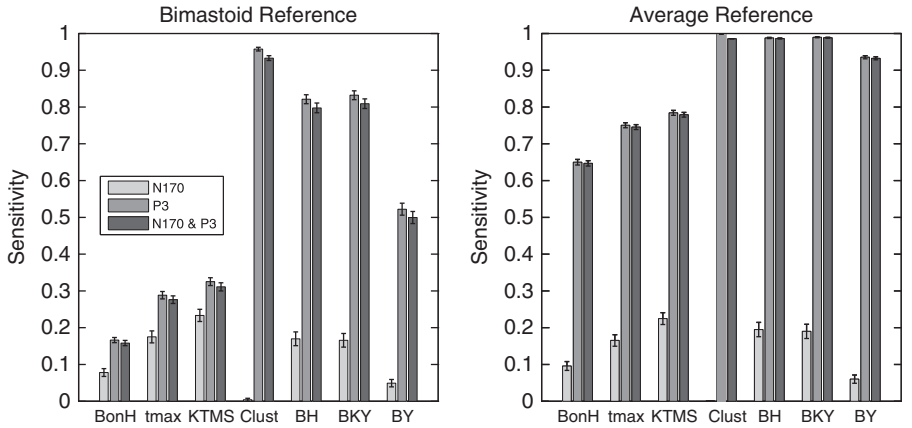

Figure 7. The sensitivity (i.e., mean proportion of time points and electrodes whose voltages differ from zero correctly detected) of seven different methods of multiple comparison correction applied to three different simulated ERP effects. Error bars indicate 95% confidence intervals.

Lastly, the choice of reference has no considerable effect on the permissiveness or relative power of the multiple comparison procedures. This is most notable for the BH and BKY procedures, since they are not guaranteed to work when some pairs of tests are negatively correlated, and about half of all test pairs are negatively correlated when the average reference is used (see Simulation Parameters). Some readers might note that all the test procedures are more powerful when applied to the average reference data than the bimastoid-referenced data. This is because the magnitude of the background noise is greatly reduced when the average reference is used but the magnitude of the effects is the same for both references. Keep in mind that this is an artifact of the simulation since effects were added after the background noise was converted to average reference in order to keep the number of false null hypotheses constant across the two types of reference. With real data, effects like the ones used here that appear at largely or entirely a single polarity at utilized sensors will also have their magnitude diminished when converting to the average reference. Thus, these results should not be taken as evidence that the average reference is generally more powerful than the bimastoid reference.

Discussion

The purpose of the simulation studies presented in this article was to evaluate the accuracy and relative power and permissiveness of several popular multiple comparison correction procedures as applied to mass univariate analysis of ERP/ERF data. These procedures differ primarily in the type of false discovery statistic each explicitly controls, which makes the methods differ in their ability to detect various types of effects, their likelihood to make false discoveries, and the certainty they provide regarding the statistical significance of any single test result. Moreover, some

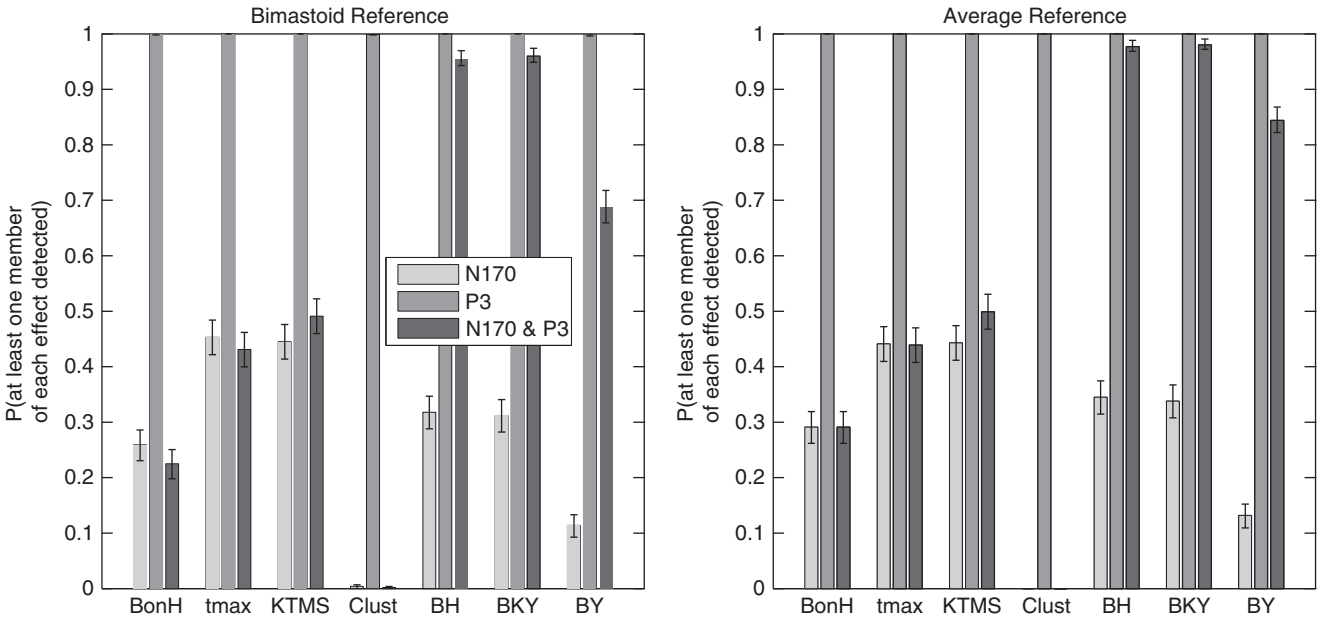

Figure 8. The probability that at least one element (i.e., a difference from zero at one time point and electrode) of each effect in the data (i.e., N170 and/or P3) is detected by seven different methods of multiple comparison correction applied to four different simulated ERP effects. Error bars indicate 95% confidence intervals.

methods are not generally guaranteed to work or may be undesirably sensitive to some deviations from the null hypothesis (e.g., differences in variance between populations) that are not of interest. Consequently, simulation studies like those reported here are critical for helping researchers know which multiple comparison correction procedures are best suited for a particular research question and how much one can trust the results of a particular procedure.

Sensitivity of Independent Samples Permutation Tests to Differences in Population Variance

The first issue we addressed in this paper was the sensitivity of an independent samples permutation test to differences in population variance. Independent samples permutation tests, like the most popular parametric independent samples tests (i.e., *t* test and analysis of variance [ANOVA]), are based upon the null hypothesis that the two samples come from identically distributed populations. Thus, the test may be sensitive to any difference between populations (e.g., differences in variance), not just differences in their means. Since such differences may reflect different amounts of noise in the two samples that are not generally of interest to ERP/ERF studies, it is important to know how likely the procedure is to detect such differences.

Our simulations, like those of Murphy (1967), show that a permutation test based on the standard independent samples *t* statistic is rather insensitive to differences in population variance when the samples being compared are of equal size. Critically, however, when the samples differ in size by even a small amount, the test results can be considerably affected by differences in variance (e.g., a FWER around 20% when the nominal FWER is 5%). This sensitivity increases as the difference in sample size or variance between groups grows. This sensitivity can be reduced by using alternative test statistics: for example, Welsh's *t* or t_{dif}. This is especially true of t_{dif}, which we find to be remarkably insensitive to differences in variance. At the same time, however, these alternative test statistics are also less sensitive to differences in means than the standard *t* statistic. Consequently, the best approach to applying independent samples permutation tests is to design the experiment such that sample sizes are equal and to use the standard *t* statistic. When this is not possible, it is essential to compare the variances of the two populations being compared. If there is evidence (or strong *a priori* reason to believe) that the two populations differ in variance, one should use Welsh's *t* if the difference in variance and sample sizes is small (e.g., a between-group sample size ratio of 1.25 and variance ratio of 1.4) and t_{dif} if those differences are larger. Note that, although we did not investigate the sensitivity of cluster-based permutation tests or tests based on the independent samples *F* statistic, we see no reason why these results would not be generally true of those procedures as well.

Accuracy, Permissiveness, and Power of FDR Methods

Of all the methods investigated here, the FDR control methods developed by Benjamini and colleagues perhaps come with the greatest uncertainty since two of the methods, BH and BKY, are not generally guaranteed to work, and any FDR method could, in principle, produce an alarmingly high proportion of false discoveries with some frequency (Korn et al., 2004). However, like Hemmelmann et al. (2005), our study found no evidence of either of these possible problems using realistic ERP data, a conventional FDR level of 5%, and a realistic number of comparisons. With higher FDR levels (e.g., 10%—Korn et al., 2004), high proportions of false discoveries might occur with much greater frequency. Thus, such FDR levels should probably be avoided. With fewer comparisons than what we've used here, it is also possible that the BH and BKY methods would not accurately control FDR (Clarke & Hall, 2009). However, given that we know of no cases in which either of these methods has terribly failed at controlling FDR,[7] we suspect that these methods will generally perform rather accurately with ERP/ERF data regardless of the number of comparisons. Researchers concerned about their accuracy can readily perform simulations like those reported here to assess their performance for other types of comparisons or use the BY procedure, which is always guaranteed to control FDR.

Regarding power, the BH and BKY procedures appear to be the most generally useful since they have relatively good power for detecting both very narrowly and broadly distributed effects (see also Hemmelmann et al., 2005; Lage-Castellanos et al., 2010). Thus, these procedures are probably the best suited for mass univariate analysis of ERP/ERF data in general unless one needs to be certain as to the reliability of every single test result (i.e., one needs strong control of FWER) or one is primarily interested in broadly distributed effects.

Permissiveness and Power of Cluster-Based Permutation Tests with Weak Control of FWER

Since cluster-based permutation tests provide only weak control of FWER, it is not clear how likely they are to falsely declare individual tests significant when some null hypotheses are actually false (i.e., there is one or more effect in the data). Our simulations found that the permissiveness of the maximum cluster-level mass statistic is comparable to or better than the other methods investigated here when this is the case. Moreover, like Maris and Oostenveld (2007), we found that the cluster-based test was the best of all the procedures we compared at detecting broadly distributed effects. This power, though, comes at the cost of a pronounced insensitivity to very focally distributed effects. Thus, the cluster-based test is probably the best suited for mass univariate analysis of ERP/ERF data when one is not interested in potentially very focally distributed effects, unless one needs to be certain as to the reliability of every single test result (i.e., one needs strong control of FWER).

Permissiveness and Power of Permutation-Based Strong Control of FWER and GFWER

Of the methods studied herein, the procedures that provide the greatest certainty as to the significance of any single test result are permutation-based control of FWER and GFWER. Again, strong control of FWER provides the same degree of certainty that any single test result is significant as Bonferroni correction or an *a priori* selective test. GFWER control provides somewhat less

7. The worst potential failure of FDR control (by any of the FDR methods studied here) that we know of is the aforementioned simulation study by Lage-Castellanos et al., which found that the BH procedure produced an FDR of 9.7% when the nominal rate was 5%. As already mentioned, though, (see Previous Work on FDR Permissiveness), it is impossible to judge the accuracy of this result.

certainty since it allows up to a certain number of false discoveries with a specified likelihood. This greater uncertainty of GFWER control can be problematic when the number of truly false null hypotheses is small relative to the number of allowed false discoveries as it will result in a high proportion of false discoveries (e.g., Figures 4 & 6).

Like Hemmelmann et al. (2005), we found that permutation-test based strong control of FWER and GFWER proved to be the most powerful methods when only a very small proportion of null hypotheses are false. This is due to the fact that these methods exploit the correlations between tests (unlike FDR procedures and Bonferroni correction) and are not biased towards detecting only broadly distributed effects (unlike cluster-based tests). The GFWER procedure, with one false discovery allowed, proved to be only slightly more sensitive than strong FWER control. It is not clear if this increased power is worth the extra uncertainty of the GFWER procedure. However, if the GFWER procedure were allowed to make more false discoveries (e.g., 2—Hemmelmann et al., 2005), its relative increase in power would improve. In light of this, the strong FWER permutation test procedure is probably the best method to use when one needs to be certain of the reliability of any single test result (e.g., assessing the lower bounds on effect onset—Groppe et al., 2010; Hillyard et al., 1973; Johnson & Olshausen, 2003), or one suspects that only a small proportion of null hypotheses is false. GFWER control is most useful when one has some idea as to how many null hypotheses to expect and strong FWER control is unlikely to be sufficiently powerful.

Conclusion

Given their minimal *a priori* assumptions and high temporal and spatial resolution, mass univariate analysis is a useful addition to the standard statistics toolbox of the ERP/ERF methodology. A key parameter of a mass univariate analysis is the choice of multiple comparison correction procedure. The simulation results reported here suggest that all of the six popular methods we investigated perform accurately when applied to conventional ERP/ERF analysis and are not prone to alarmingly high proportions of false discoveries. Thus, these procedures appear to be generally valid for ERP/ERF analysis though some care must be taken when applying them to independent samples comparisons to avoid mistaking differences in variance for differences in means.

These results also illustrate various trade-offs that the different methods make as to the type of effects they are best suited to detect and the degree of certainty they provide as to the significance of any single test result. The circumstances in which these procedures appear to be best suited are summarized in Table 2 of the companion article to this paper (Groppe et al., 2011). These simulation results will help investigators understand which method best fits their particular needs and to better interpret such analyses. If researchers would like additional background and explanation on how these procedures work, the companion article provides a tutorial review of these methods (Groppe et al., 2011). In addition, to help researchers apply these methods to their own data, we have provided freely available MATLAB software, called the "Mass Univariate ERP Toolbox." The software is compatible with the EEGLAB toolbox (Delorme & Makeig, 2004) as well as the ERPLAB toolbox (http://erpinfo.org/erplab). The software, software documentation, and a tutorial are available on the toolbox wiki (http://openwetware.org/wiki/Mass_Univariate_ERP_Toolbox). The EEGLAB toolbox and FieldTrip (Oostenveld, Fries, Maris, & Schoffelen, 2011) MATLAB software packages implement some of these procedures as well.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*, 289–300.

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

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*, 1165–1188.

Bentin, S., Mouchetant-Rostaing, Y., Giard, M. H., Echallier, J. F., & Pernier, J. (1999). ERP manifestations of processing printed words at different psycholinguistic levels: Time course and scalp distribution. *Journal of Cognitive Neuroscience*, *11*, 235–260. doi: 10.1162/089892999563373.

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

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

Clarke, S., & Hall, P. (2009). Robustness of multiple testing procedures against dependence. *The Annals of Statistics*, *37*, 332–358. doi: 10.1214/07-AOS557.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*, 9–21. doi: 10.1016/j.jneumeth.2003.10.009.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of American Statistical Association*, *99*, 96–104. doi: 10.1198/016214504000000089.

Groppe, D. M., Choi, M., Topkins, B., & Kutas, M. (2009). An event-related potential investigation of the phonemic restoration effect. *Cognitive Neuroscience Society Annual Meeting Program,* San Francisco, CA. 156.

Groppe, D. M., Choi, M., Huang, T., Schilz, J., Topkins, B., Urbach, T. P., & Kutas, M. (2010). The phonemic restoration effect reveals pre-N400 effect of supportive sentence context in speech perception. *Brain Research*, *1361*, 54–66. doi: 10.1016/j.brainres.2010.09.003.

Groppe, D. M., Urbach, T. U., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*. doi: 10.1111/j.1469-8986.2011.01273.x.

Hemmelmann, C., Horn, M., Reiterer, S., Schack, B., Susse, T., & Weiss, S. (2004). Multivariate tests for the evaluation of high-dimensional EEG data. *Journal of Neuroscience Methods*, *139*, 111–120. doi: 10.1016/j.jneumeth.2004.04.013.

Hemmelmann, C., Horn, M., Susse, T., Vollandt, R., & Weiss, S. (2005). New concepts of multiple tests and their use for evaluating high-dimensional EEG data. *Journal of Neuroscience Methods*, *142*, 209–217. doi: 10.1016/j.jneumeth.2004.08.008.

Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science*, *182*, 177–180. doi: 10.1126/science.182.4108.177.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

Johnson, J. S., & Olshausen, B. A. (2003). Timecourse of neural signatures of object recognition. *Journal of Vision*, *3*, 499–512, doi: 10:1167/3.7.4.

Kim, K. I., & van de Wiel, M. A. (2008). Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics*, *9*, 114. doi: 10.1186/1471-2105-9-114.

Korn, E. L., Troendle, J. F., McShane, L. M., & Simon, R. (2004). Controlling the number of false discoveries: Application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, *124*, 379–398. doi: 10.1016/S0378-3758(03)00211-8.

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

Lee, T. W., Girolami, M., & Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation*, *11*, 417–441, doi: 10.1162/089976699300016719.

Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, MA: MIT Press.

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

Murphy, B. P. (1967). Some two-sample tests when the variances are unequal: A simulation study. *Biometrika*, *54*, 679–683.

Naatanen, R., & Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. *Psychological Bulletin*, *125*, 826–859. doi: 10.1037/0033-2909.125.6.826.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience, 2011*, 156869. doi: 10.1155/2011/156869

(Received March 14, 2011; Accepted June 19, 2011)