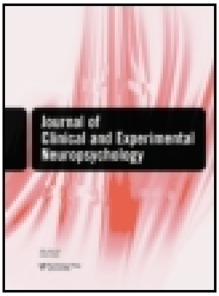

Journal of Clinical and Experimental Neuropsychology

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ncen20

Enhancing Power While Controlling Family-Wise Error: An Illustration of the Issues Using Electrocortical Studies

Paul J. Yoder [a] , Jennifer Urbano Blackford [a] , Niels G. Waller [a] & Geunyoung Kim [a]

[a] Kennedy Center, Vanderbilt University Nashville TN USA

Published online: 09 Aug 2010.

To cite this article: Paul J. Yoder , Jennifer Urbano Blackford , Niels G. Waller & Geunyoung Kim (2004) Enhancing Power While Controlling Family-Wise Error: An Illustration of the Issues Using Electrocortical Studies, Journal of Clinical and Experimental Neuropsychology, 26:3, 320-331, DOI: 10.1080/13803390490510040

To link to this article: http://dx.doi.org/10.1080/13803390490510040

Journal of Clinical and Experimental Neuropsychology
2004, Vol. 26, No. 3, pp. 320–331

Psychology Press
Taylor & Francis Group

Enhancing Power While Controlling Family-Wise Error: An Illustration of the Issues Using Electrocortical Studies

Paul J. Yoder, Jennifer Urbano Blackford, Niels G. Waller, and Geunyoung Kim
Kennedy Center, Vanderbilt University, Nashville, TN, USA

ABSTRACT

This study examined the relative family-wise error (FWE) rate and statistical power of multivariate permutation tests (MPTs), Bonferroni-adjusted alpha, and uncorrected-alpha tests of significance for bivariate associations. Although there are many previous applications of MPTs, this is the first to apply it to testing bivariate associations. Electrocortical studies were selected as an example class because the sample sizes that are typical of electrocortical studies published in 2001 and 2002 are small and their multiple significance tests are typically nonindependent. Because Bonferroni adjustments assume independent predictors, we expected that MPTs would be more powerful than the Bonferroni adjustment. Results support the following conclusions: (a) failure to control for multiple significance testing results in unacceptable FWE rates, (b) the FWE rate for the MPTs approximated the alpha set for the analyses, and (c) the statistical power advantage that MPTs provide over Bonferroni adjustments is important when using small sample sizes such as those that are typical of recent electrocortical studies.

In this paper, we argue that the extent to which some fields of inquiry produce replicable results is limited by the common use of small sample sizes and the lack of control for multiple significance testing. The result is elevated Type I and Type II error rates. Further, we argue that there is a need for statistical methods that maximize statistical power while controlling family-wise error (FWE) rate. We define FWE rate as the probability of at least one false positive finding out of a set of significance tests directed at addressing the same research question. The primary purpose of this paper is to describe an application of multivariate permutation tests (MPTs) to examine multiple bivariate associations that controls FWE rate. Although many applications of MPTs have been discussed in previous literature (Blair & Karniski, 1994; Nichols & Holmes, 2001; Pesarin, 2001; Westfall & Young, 1993), this paper is the first to describe its application to testing bivariate associations. We also estimate the relative FWE rate and statistical power of this and two other methods of testing the significance of bivariate associations.

Although many scientific fields would benefit from methods of controlling FWE rate while maximizing statistical power, we present these issues within the context of electrocortical studies as an example class of studies. We have selected the electrocortical area because the characteristics of such studies make the most common methods of protecting the FWE rate for testing multiple bivariate associations inappropriate. By ‘‘electrocortical’’ studies, we mean studies using event-related potential (ERP) and electroencephalography (EEG) to measure brain activity. One

Address correspondence to: Paul J. Yoder, Vanderbilt University, Peabody Box 328, Nashville, TN 37203, USA.
E-mail: Paul.Yoder@vanderbilt.edu

Accepted for publication: May 15, 2003.

10.1080/13803390490510040$16.00 © Taylor & Francis Ltd.

of the common characteristics of electrocortical studies is that the number of variables exceeds sample size. Additionally, significance tests are often nonindependent in electrocortical studies. Many of the variables in electrocortical studies are measures of EEG activity at various electrodes and various poststimulus latencies. These variables are often highly correlated because of the spatial proximity of electrodes and the temporal proximity of ERP time samples or components. Later in the paper, we explain why these attributes are problematic for two popular approaches to protecting FWE rate. Although we draw our examples from the electrocortical literature, scientists in any field who design studies with more variables than subjects and with multiple nonindependent significance tests are likely to find the present simulation study useful.

Probable Obstacles to Replication in the Electrocortical Field

The bias against publishing nonsignificant results makes it extremely difficult to estimate the proportion of electrocortical results that are replicated at a specific level of analysis (i.e., the level of the electrode, component, and ERP quantification method; Greenwald, 1993). Unless specific procedures are used to protect FWE rate, the use of multiple significance testing to examine a single research question increases the probability of detecting sample-specific results (i.e., increased Type I error rates; Westfall & Young, 1993). Small sample sizes reduce statistical power and increase Type II error rates (Cohen, 1988). Therefore, elevated Type I and Type II error rates are likely to the extent that electrocortical studies use small sample sizes and multiple significance testing without protection against inflated FWE rates. Both types of errors reduce the probability of replication. Most scientists would agree that sample-specific results do not warrant publishing. At best, sample-specific results may direct other researchers in unproductive directions. At worst, unreplicated, sample-specific results may motivate incorrect clinical decisions (e.g., over-diagnosis).

The statistical power issue is particularly salient for the electrocortical field because the sample sizes of *most* such studies are relatively small. A *PsycInfo* review of EEG and ERP studies in 2001 and 2002 yielded 241 studies. Figure 1 provides the histogram for the sample sizes of these studies. It should be noted that the distribution is positively skewed. The mode, median, and mean sample sizes were 15, 20, and 34, respectively. Excluding two outliers (i.e., sample sizes that were greater than 3 *SD* from the

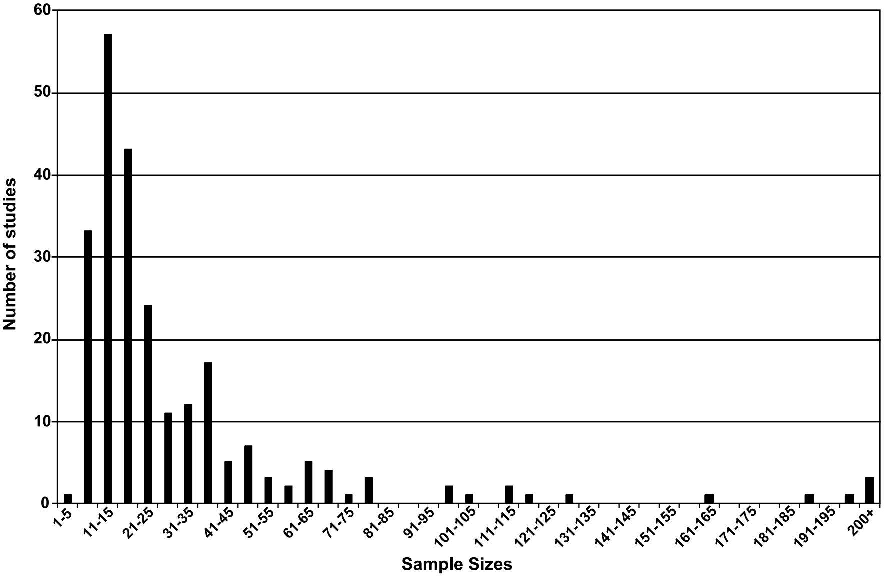

Fig. 1. Histogram of the sample sizes of 241 ERP studies published in 2001 and 2002.

mean), the mean sample size was 29 ($SD = 30$). With the mode sample size (i.e., 15), power analyses to detect bivariate associations (i.e., Pearson Product Moment correlations) of .1, .3, and .5 yielded statistical power estimates of 7%, 20%, and 54%, respectively. To achieve 80% power to detect a single association with 15 subjects, the Pearson's r must be at least .61.

The other attribute of many electrocortical studies that contributes to a low replication rate is the use of multiple significance tests to address a single research question without protection against inflated FWE rates. It has been shown repeatedly that the probability of incorrectly rejecting a null hypothesis increases as the number of statistical tests examining the same hypothesis increases (Hochberg & Tamhane, 1987; Miller, 1981). We call this the "multiple testing problem."

There are many examples of multiple significance testing in the electrocortical literature. Electrocortical scientists must deal with relatively vague theories, incomplete knowledge, multiple electrodes, and multiple poststimulus latencies of interest. It is not unusual to see studies that use 128 electrodes (e.g., Potts, Dien, Hartry-Speiser, McDougal, & Tucker, 1998). Several methods are used to explore multiple poststimulus latencies. For example, some studies (e.g., Kramer, Trejo, & Humphrey, 1995) examine multiple ERP components (e.g., N1, P2, N2, P3). Others (e.g., van Laar et al., 2002) analyze successive time intervals that contain several time samples to cover a large portion of the ERP wave (e.g., mean voltage in 15 successive 50 ms intervals). Others (e.g., Trainor, Samuel, Despardins, & Sonnadara, 2001) use multiple significance tests for adjacent time samples (i.e., a voltage reading every 4 ms) to identify the poststimulus latency at which amplitudes are different from zero. It should be noted that the problem is not the use of multiple significance testing, but the use of such without protection against inflated FWE rates.

Current Methods of Controlling for Multiple Significance Testing in Electrocortical Studies

The most common method of addressing research questions about multiple latencies and electrodes is to use repeated measures ANOVA and treat latencies and electrodes as within-subject factors (Cohen, 1987). For example, Potts et al. (1998) used a task (2) × stimulus (2) × site (29) × hemisphere (2) repeated measures ANOVA to examine whether neural responses vary as a function of task-relevant auditory stimuli. Repeated measures ANOVA assumes that covariances among the levels in a within-subjects factor are equal (Huynh & Feldt, 1970). In a common application of repeated measures ANOVA in the electrocortical literature, electrodes are the levels in a within-subject factor called "site" (e.g., Potts et al., 1998). Because electrodes that are proximal are more highly correlated than electrodes that are distal from each other, this assumption is frequently violated in electrocortical data (Vasey & Thayer, 1987). Using the well-known Greenhouse and Geisser (1959) correction for violations of this assumption results in reduced statistical power (Karniski, Blair, & Snider, 1994) and only approximates the F-distribution (Vasey & Thayer, 1987).

Another parametric approach to multiple significance testing that does not assume equal covariances among levels in a within-subject factor is MANOVA. This application of MANOVA treats levels of within-subject factors as multiple dependent variables (McCall & Appelbaum, 1973). For example, one might treat electrodes or multiple poststimulus latencies as dependent variables and manipulated conditions as an independent variable (Vasey & Thayer, 1987). If the omnibus test is significant, post hoc tests that control FWE rate can be used to identify univariate effects (Hochberg & Tamhane, 1987). Unfortunately, this approach produces invalid results when the number of variables exceed the sample size, which is common in the electrocortical literature (Galan, Biscay, Rodriguez, Perez-Abalo, & Rodriguez, 1997).

Both of these approaches are designed for tests of the mean difference among groups or conditions. However, many research questions in ERP studies involve testing an association between individual differences on the ERP variables and individual differences on some behavioral measure. For example, Dawson, Finley, Phillips, and Lewy (1989) examined the association between N1 latencies in the left hemisphere and language scores. It is inadvisable to reduce continuous variables to categorical ones so that familiar analysis methods such as ANOVA or MANOVA can be used to examine brain-behavior associations because doing so results in loss of information and reduced statistical power (Cohen, 1983). Although not commonly used in the electrocortical literature, canonical correlation is the analogous analysis method to MANOVA that allows retention of the continuous form of multiple predictors and criterion variables (Tabachnick & Fidell, 1996). Unlike the follow-up tests that control for multiple significance testing that are recommended for use after significant omnibus MANOVA tests (Hochberg & Tamhane, 1987), there are no commonly used post hoc procedures used after significant canonical correlation tests that control for multiple significance testing. The most common analytic method used to identify significant bivariate associations after significant canonical correlation tests is a series of *t*-tests with unadjusted or adjusted alpha.

If the research question requires testing significance of bivariate associations and one wants to protect FWE against inflation, one must adjust the alpha level. The most common method of adjusting alpha is to divide it by the number of significance tests (i.e., Bonferroni's correction; Bonferroni, 1950; Dunn, 1959). In the ERP literature, Kramer et al. (1995) provides an example of using this method to examine the relationship between task relevancy and ERP activity at several poststimulus latencies. The Bonferroni correction assumes that the variables in the different significance tests are independent (Blair & Karniski, 1994). For example, if an investigator studying the ERP-language association uses the Bonferroni correction, the investigator implicitly assumes that the data from the various electrodes are uncorrelated. As mentioned earlier, the brain activity measured by adjacent electrodes is frequently highly correlated (Blair & Karniski, 1994). Under these conditions, the Bonferroni-adjusted alpha will be lower than the intended family-wise alpha. That is, the Bonferroni method over-corrects the alpha and results in a loss of power when used in the context of studies with highly correlated predictors. This loss of power can be illustrated by imagining that the electrodes on the skull are perfectly correlated. A significance test of the association between data from one of these electrodes with the behavioral measure would be equivalent to a significance test of the associations between the data from all electrodes and the behavioral measure. Even in this extreme case, the Bonferroni adjustment still divides the alpha by the number of significance tests.

Nonparametric Approaches to Controlling for Multiple Significance Testing

All of the above methods of controlling for multiple significance testing use parametric methods for estimating the *p* value of test statistics. Most common parametric methods assume multivariate normality (Tabachnick & Fidell, 1996). Electrocortical data often violate this assumption (Faux & McCauley, 1990; Karniski et al., 1994). Several scientists in the neuro-imaging field have suggested using nonparametric methods to protect against inflated FWE rate, in part, because such methods do not assume multivariate normality (Blair & Karniski, 1994; Faux & McCauley, 1990; Galan et al., 1997; Nichols & Holmes, 2001).

Multivariate permutation testing (MPT) and multivariate bootstrapping operations are two such nonparametric approaches (Westfall & Young, 1993). Westfall and Young (1993) lay out the logical considerations for choosing between permutation tests and bootstrapping operations to control for multiple significance testing. Although we recognize that there are differing opinions regarding this general choice (Westfall & Young, 1993 for review), many neuro-imaging scientists have elected to use permutation tests over bootstrapping (e.g., Blair & Karniski, 1994; Faux & McCauley, 1990; Galan et al., 1997; Nichols & Holmes, 2001). This preference may, in part, be based on the fact that permutation tests assume that the probability values are derived from random processes within the selected sample only. Bootstrapping, strictly speaking, assumes that probability values are derived from random samples of a larger population. Given that electrocortical studies virtually never use random sampling of a population, it is most appropriate for such scientists to use a testing method that does not assume random sampling.

Westfall and Young (1993) and Pesarin (2001) describe the general multivariate permutation testing (MPT) procedure. Blair and Karniski (1994) and Nichols and Holmes (2001) describe the procedure for *within-subject comparisons* using neuro-imaging examples. In general, the MPT process begins by computing the test statistic for all planned significance tests from the observed data. The observed maximum (max) test statistic is the highest of these test statistics. The significance of the observed max statistic is determined by comparing it to a distribution of the max test statistics that are contributed by many permutations or random shufflings (i.e., random sampling without replacement) of the observed data. Exactly which variable is randomly shuffled depends on the type of effect being tested. In a two-tailed test, the *p* value is determined by the proportion of postshuffle max test statistics with absolute values that exceed the absolute value of the observed max test statistic. If the observed maximum test statistic is significant (i.e., *p* value < alpha), then the data for that variable is discarded from the observed and all permuted data sets and the process is repeated until a nonsignificant finding occurs. The discarding of the data is what "adjusts" the distribution for multiple significance testing.

Technically, a permutation test uses all possible shufflings (i.e., permutations) of the observed data (Edgington, 1987). However, it has been shown that using very many (e.g., 10,000) data permutations produce extremely accurate approximations of the exact probability value (i.e., that produced by all possible permutations of the data; Edgington, 1969). With relatively few variables or with variables that are highly correlated, MPTs can be accurate with as few as 1,000 permutations (Edgington, 1969). When many shufflings are used instead of all possible shufflings, the approach is called a "sampled permutation test" (i.e., approximate randomization tests). Most people use the sampled permutation test because the number of possible permutations with studies that have even 10 subjects is quite large (i.e., $10! = 3{,}628{,}800$).

Blair and Karniski (1994) tested the family-wise error for a 2-level within-subjects comparison using MPTs and Bonferroni adjusted alpha. Using "moderately correlated data" of unspecified magnitude and an intended family-wise alpha of .05, they found that the actual family-wise error for MPT and the Bonferroni-adjusted alpha method was .05 and .02, respectively. In the same paper, the authors demonstrated that the sample and effect sizes influence statistical power of MPT in the usual way (i.e., more is better). In a separate paper, Blair, Higgins, Karniski, and Kromrey (1994) found that statistical power of MPT remained strong even when the number of variables exceeded the number of subjects. This latter finding is particularly relevant to electrocortical studies because the number of variables frequently exceeds sample size.

The MPT approach to testing significance of multiple bivariate associations has not been described in either of the major texts (Pesarin, 2001; Westfall & Young, 1993) or in the current tutorials on MPT in the neuro-imaging literature (Blair & Karniski, 1994; Nichols & Holmes, 2001). To the authors' knowledge, the first application of the MPT approach to testing multiple bivariate associations reported in the electrocortical literature was reported in Henderson, Yoder, Yale, and McDuffie (2002) and conducted in our lab. However, this report did not describe the method in detail or indicate the FWE rate or statistical power of this application of MPT. Additionally, the relative FWE rate and statistical power of MPT and Bonferroni-adjusted alpha has not been compared when testing the significance of bivariate associations. The present paper seeks to address these needs.

The Mechanics of MPT of Bivariate Correlations

Figure 2 lists the steps that are used to implement the MPT approach to testing the significance of a series of bivariate associations. In general, the bivariate association indices are ranked from largest to smallest. The significance of the largest of these associations (i.e., max *r*) is tested first. To model random associations between predictors and the criterion variable, the values of the criterion variable are shuffled across subjects. This process results in participants retaining their observed values on the predictor variables. In this way the intercorrelation of predictors is

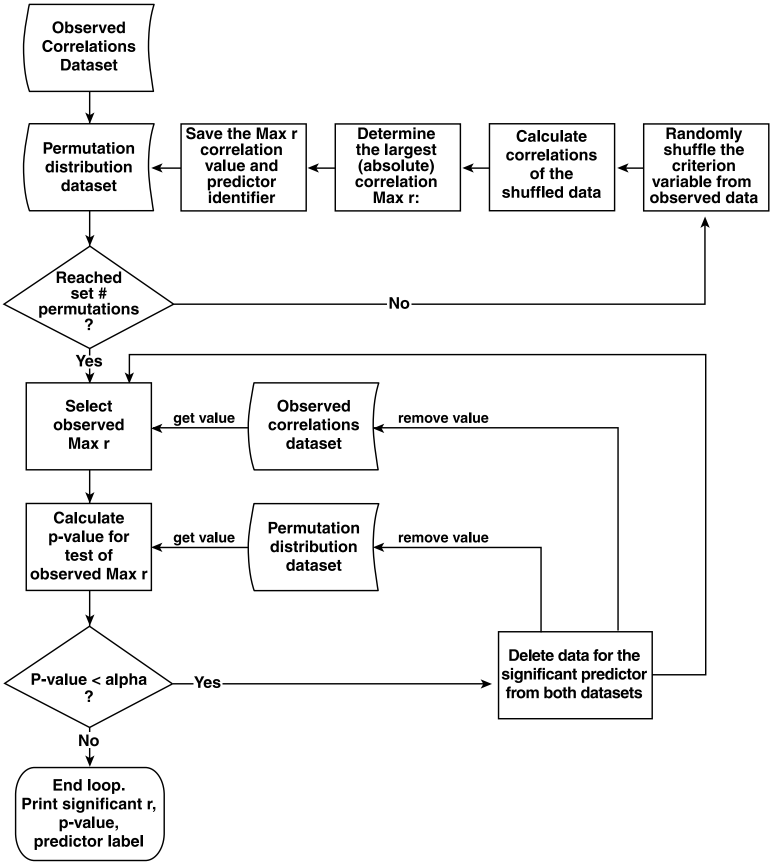
Observed Correlations Dataset
Permutation distribution dataset
Save the Max r correlation value and predictor identifier
Determine the largest (absolute) correlation Max r:
Calculate correlations of the shuffled data
Randomly shuffle the criterion variable from observed data
Reached set # permutations ?
No
Yes
Select observed Max r
get value
Observed correlations dataset
remove value
Calculate p-value for test of observed Max r
get value
Permutation distribution dataset
remove value
P-value < alpha ?
Yes
Delete data for the significant predictor from both datasets
No
End loop. Print significant r, p-value, predictor label

Fig. 2. General steps used by the multivariate permutation test of significance of bivariate associations.

reflected in the postshuffle max *r* coefficients. The probability distribution used to test the significance of the observed max *r* is composed of one max *r* value from each of the postshuffle data sets. If the observed max *r* is found significant, the postshuffle *r* values for the association between the significant predictor and the criterion variable are deleted from the probability distribution.

To ensure equal sensitivity across predictors, it is necessary to compute a statistic that standardizes the effect for differing variances across electrodes and latencies (Nichols & Holmes, 2001). Therefore, the predictor-criterion covariances are standardized by both variables' variance estimates (i.e., Pearson's *r* is computed). It is not necessary to compute the corresponding *t* statistic because the probability of *r* will be tested against the aforementioned empirical probability distribution, not the familiar *t* distribution.

Figure 3 presents example observed and postshuffle data, and the corresponding *r* values. Although using only three participants is inadvisable, we provide this example with 3 participants so that the entire analytic process can be easily followed. However, it is inadvisable to compute correlation coefficients on samples under 10 participants. In fact, at least 4 subjects are needed to produce *p* values between 0 and .05. In a two-tailed significance test, the absolute value of the *r* coefficients are ranked to identify the max $|r|$. In our example, the observed max $|r|$ was .91. The distribution of max $|r|$ values against which the observed max $|r|$ is tested is as follows: .91, .82, 1.0, .66, .96, and .99. It should be noted that in an exact test, the observed value is included in the probability distribution. Three of the postshuffled max $|r|$ values were greater than the observed. Therefore, the probability that $|.91|$

id	criterion	Predictor1	Predictor2	Predictor3
1	1	5.25	4.94	5.48
2	7	5.14	4.73	5.20
3	6	5.08	4.91	5.09
Pearson's *r*		-.87	-.73	-.91*
Shuffle 1				
1	7	5.25	4.94	5.48
2	6	5.14	4.73	5.20
3	1	5.08	4.91	5.09
Pearson's *r*		.80	-.23	.82*
Shuffle 2				
1	7	5.25	4.94	5.48
2	1	5.14	4.73	5.20
3	6	5.08	4.91	5.09
Pearson's *r*		.32	1.00*	.39
Shuffle 3				
1	6	5.25	4.94	5.48
2	7	5.14	4.73	5.20
3	1	5.08	4.91	5.09
Pearson's *r*		.66*	-.52	.60
Shuffle 4				
1	6	5.25	4.94	5.48
2	1	5.14	4.73	5.20
3	7	5.08	4.91	5.09
Pearson's *r*		.01	.96*	.09
Shuffle 5				
1	1	5.25	4.94	5.48
2	6	5.14	4.73	5.20
3	7	5.08	4.91	5.09
Pearson's *r*		-.98	-.48	-.99*

* denotes max $|r|$

Fig. 3. Example of MPT process for testing significance of association between a criterion and three predictors with three subjects' data.

occurred by chance was 3/6 or .5. Because .5 is not less than .05, the process stops here. If the probability value had been below .05, we would have eliminated the data for predictor number three from all data sets and repeated the process until the max $|r|$ was not significant.

In the remainder of this paper, we describe a Monte Carlo study that was designed to estimate the FWE rate of MPTs for bivariate associations. The FWE and statistical power of the Bonferroni-adjusted alpha and unadjusted-alpha methods were also computed because these are common approaches to multiple significance testing of bivariate associations in the electrocortical literature. This comparison was conducted with three sample sizes that are typical of the sample sizes of recent electrocortical studies.

METHODS

Three sample sizes were simulated based on the mean ($n = 35$) and mode ($n = 15$) of our sample of 241 electrocortical studies that were published in 2001 and 2002. The mid-point between the mean and mode sample size was selected as a third sample size condition ($n = 25$).

Five thousand experiments were generated for each sample size condition. Simulations conducted in our lab using 50,000 experiments found very similar results to those reported below. Therefore, we decided to use 5,000 to allow using CPU time to generate simulated data for three sample sizes. Other simulations that estimated Type I error rates and statistical power for various multiple comparison techniques have also used 5,000 experiments (Kromrey & LaRocca, 1995). In each data set, there was one criterion variable and 32 predictors. This number of predictors was chosen because 32 is a compromise between the recently developed high density arrays (e.g., 128 electrodes) and older electrode arrays (e.g., 16 electrodes). The predictors were divided into two sets: those designated as correlated with the criterion variable (i.e., four target predictors) and those designated as uncorrelated with the criterion variable (i.e., 28 nontarget predictors).

There is a necessary tension between two goals of our simulation. First, we wanted to use the same data sets to estimate both FWE rate and statistical power because a real trade off exists between maximizing power and minimizing FWE. If we used a separate set of data to estimate FWE rate from that used to estimate power, we might inadvertently produce unrealistic correlational structures in our simulated data sets. Second, we wanted to design the computer program that generated the simulated data sets on real electrocortical data to improve the probability that the correlational structure of the simulated data sets would be realistic (Micceri, 1989).

Therefore, the intercorrelations among the predictors in the simulated data sets were set to those occurring in the data reported in Henderson et al. (2002). One reason for modeling the simulated data after the parameters in the study by Henderson et al. was that only large effect sizes are detectable in studies using proper FWE protection with small sample sizes. The effect size of the target electrodes with the behavioral criterion variable averaged −.60. The EEG data was the average squared amplitude of brain activity from the 4–6 Hz frequency band. The rest of the correlations in the program generating the simulated data sets were based on those in the study by Henderson et al. to the extent possible within following three constraints. The first constraint was that the associations between the nontarget predictors and the criterion variable were all set to zero to allow estimation of FWE rate. The second constraint was that the associations of the four target predictors with the criterion variable were all set to the average of the four significant EEG-criterion correlations in the study by Henderson et al. (i.e., −.60). This was done to allow parsimonious reporting of statistical power. The third constraint was that the EEG data for three nontarget predictors surrounding the target predictors in the Henderson et al., study were not used to guide the program that would generate simulated data sets. The latter action was taken because these three EEG predictors were highly correlated with the target predictors and were moderately, but nonsignificantly, associated with the criterion variable in the Henderson et al. study. Because all nontarget predictors' association with the criterion variable had to be set to zero, nontarget predictors could not be highly correlated with target predictors, whose association with criterion variable was set to −.60. To replace the values for these three nontarget variables, we randomly selected the correlations of three other nontarget variables with the criterion variable to bring the number of predictors to 32, a common number of electrodes in an ERP experiment.

When estimating FWE rate and power from these simulated data sets, we used two-tailed significance tests using an alpha of .05. With 32 Pearson Product Moment correlations tested, the Bonferroni-adjusted alpha for individual significance tests was .00156 (i.e., .05/32). The MPTs were run using 1,000 permutations per simulated data set. One thousand permutations generally produces sufficiently accurate *p* values to estimate relative FWE rates and statistical power when the number of predictors are limited and the correlation among predictors is high (Nichols & Holmes, 2001).

RESULTS

Table 1 presents the descriptive statistics on the distribution of correlations in the simulated data sets for the five sets of correlations (e.g., target-target, target-criterion, etc.). It should be noted that the magnitude of the average target-criterion association was approximately −.60, while the magnitude of the average nontarget-criterion associations was approximately zero. These data indicate that the computer program accurately generated a simulated population with the known large and null effects we intended to model. It should also be noted that the intercorrelation among the EEG predictors is quite high. This is important because real EEG data are highly intercorrelated. Finally, the well-known effect of sample size on sampling error is reflected in the decrease of *SD* and range of the simulated correlations as the sample size increases.

The results pertaining to FWE and statistical power are presented in Table 2. In Table 2, values in the cells are the proportion of the 5,000 experiments that are significant by the three methods of significance testing. It should be noted in Table 2 that the results regarding FWE rate are almost identical across sample sizes across all three methods of significance testing. Therefore, the contrasts that inform us about FWE rate are those across methods of significance testing. The data in Table 2 indicate that the FWE rate of the MPT is the closest to the selected alpha. In fact, using the 1,000 permutations in the

Table 1. Descriptive Statistics for the Three Sets of 5,000 Simulated Data Sets.

Sample size	Correlation type	Mean	Median	*SD*	Minimum	Maximum
15	Target-criterion	−.59	−.61	.15	−.91	.22
25		−.60	−.61	.11	−.88	−.08
35		−.60	−.60	.10	−.84	−.19
15	Nontarget-criterion	−.01	−.02	.20	−.58	.67
25		−.01	−.02	.16	−.49	.47
35		−.01	−.01	.13	−.40	.43
15	Target–target	.93	.93	.03	.73	.98
25		.93	.93	.02	.82	.97
35		.93	.93	.02	.83	.97
15	Nontarget–nontarget	.61	.61	.03	.51	.69
25		.60	.61	.02	.53	.67
35		.60	.60	.02	.54	.66
15	Target–nontarget	.53	.54	.05	.33	.69
25		.53	.53	.04	.40	.66
35		.53	.53	.03	.41	.64

Table 2. Statistical Power and Family-Wise Error of MPT, Bonferroni, and Unadjusted-Alpha Methods of Significance Testing for Bivariate Correlations.

n	Family-wise error rate[a]			Average statistical power[b]		
	MPT	Bonferroni	Unadjusted *t*	MPT	Bonferroni	Unadjusted *t*
15	.04	.02	.34	.25	.17	.72
25	.04	.02	.36	.65	.54	.94
35	.04	.02	.34	.87	.80	.99

Note. [a]Family-wise error rate = proportion of 5,000 simulated data sets in which at least one nontarget-criterion correlation was statistically significant.

[b]Average statistical power = proportion of 5,000 simulated data sets in which a particular target-criterion association was statistically significant averaged across the four target predictors.

MPTs that we implemented, it was only slightly under the nominal alpha. In contrast, the FWE rate for the Bonferroni correction was much lower than the nominal alpha, while that of the unadjusted *t*-test was much greater than the nominal alpha.

In contrast to the FWE rate, the data in Table 2 indicate that statistical power varied by significance testing method and by sample size. Looking across sample sizes, the unadjusted-alpha *t*-test method is the most powerful of the three methods. However, this power advantage is bought at the expense of an extremely high family-wise error rate. The FWE rate of the unadjusted-alpha *t*-test is much greater than most investigators would consider acceptable. Of the two methods that maintain FWE rate below the selected alpha (i.e., .05), MPT was the more powerful technique. Looking across sample sizes, the average difference in power between MPT and Bonferroni-adjusted *t*-tests was approximately 9% ($SD = 2\%$).

The power advantage of MPT over the Bonferroni-adjusted *t*-test is important for the small sample sizes that are common in electrocortical studies. For the simulation with 15 participants, all three methods had insufficient power. For the simulation with 35 participants, all three approaches provided sufficient power to detect the large effect size we modeled. For the simulation with 25 participants, the statistical power advantage of MPT over Bonferroni becomes salient.

DISCUSSION

The purpose of this article was to describe in detail the multivariate permutation test (MPT) approach to testing the significance of multiple bivariate associations. We conducted three simulations to estimate the relative power and family-wise error (FWE) rate for this approach, the Bonferroni-adjusted alpha, and the unadjusted *t*-test. In these simulations, multiple testing inflated FWE to an unacceptable degree. Therefore, the power advantage shown by unadjusted-alpha *t*-tests was offset by the unacceptably high FWE rate of these tests. The conservative ethic of the scientific method suggests that we first protect against inflated FWE rates. Both Bonferroni and MPT approaches were effective in controlling the FWE rate. However, MPT had more statistical power than the Bonferroni approach. This advantage was due to the fact that MPT takes into account the intercorrelation of the predictors, while Bonferroni assumes predictors are independent. Logic suggests that the more intercorrelated the predictors, the larger the difference between the statistical power of the MPT and Bonferroni methods of alpha adjustment.

The power advantage MPT has over the Bonferroni correction becomes salient when we consider the average sample size of recently published electrocortical studies. The apparent good news that all three methods provided sufficient statistical power to detect the effect at the average sample size for recent electrocortical studies ($n = 35$) was only true for the effect size we modeled: a very large one. This will not be so for smaller effect sizes. This is demonstrated by the fact that the Bonferroni-adjusted alpha produced exactly 80% power to detect the modeled correlation of $-.60$ with 35 participants. The bad news was that none of the methods tested were sufficiently powerful to detect even the very large effect size we modeled when the mode sample size ($n = 15$) for recent electrocortical studies was utilized.

Fortunately, MPTs are used with increasing frequency in the PET, fMRI, and LORETA literature (Nichols & Holmes, 2001). Some MPT applications with fMRI use the pooled variance from surrounding pixels (i.e., smoothing) to improve the variance estimate used to compute standardized test statistics (Holmes, Blair, Watson, & Ford, 1996). Simulation studies show that such a method improves the statistical power of MPTs (Holmes et al., 1996). Smoothing increases power by improving the reliability of the neuro-imaging variance estimate through the use of variance estimates from several sources of neuro-imaging data. An increase in reliability decreases the confidence interval of the correlation coefficient. It should be noted that the present study's FWE rate was slightly lower than the theoretical alpha (i.e., .04, not .05). This slightly smaller-than-expected FWE rate could be due to using the variance estimate at the single electrode to compute the test statistic. If we had used the weighted average of variance estimates from surrounding electrodes to standardize the covariation with the criterion variable, power and FWE accuracy may have been optimized. This is a topic for future research.

Future research is also needed to determine sufficient sample sizes in electrocortical studies when MPTs are used. Clearly, the required sample size for acceptable statistical power levels varies by effect size (Cohen, 1988). This was illustrated by the reduction in range and *SD* of the simulated correlations as sample size increased in the current study. Additionally, logic indicates that the higher the intercorrelation among predictors, the more powerful the MPT. Current electrocortical studies rarely report the intercorrelation among EEG or ERP predictors. Doing so would provide the information needed to generate more realistic simulation studies. Such simulation studies are needed to provide realistic modeling of how the strength of the intercorrelation among predictors and the effect size with the criterion variable combine to affect required sample size to achieve acceptable power levels when MPTs are used.

CONCLUSION

There are many advantages of the MPT over parametric methods of controlling for multiple significance testing in electrocortical studies. First, MPT adjusts the probability value of each test statistic by the justified amount given the intercorrelation among predictors. Second, MPT results are still valid when the number of variables exceeds the sample size. Third, MPT makes limited assumptions about the data. In contrasts between groups or between conditions, MPT only assumes that the sample variables are symmetrically distributed (Pesarin, 2001). In tests of bivariate associations, MPT only assumes that analysis units (e.g., subjects) are exchangeable under the null hypothesis (Pesarin, 2001; Westfall & Young, 1993). In this context, exchangeability means that the permutations used in the MPT do not violate what can occur in nature. For example, if the units of analysis have mixed dependence, the data are not fully exchangeable under the null hypothesis (Hayes, 1996).

When the data do not fit the assumptions of parametric analyses, MPTs can be more powerful than comparable parametric approaches (Nichols & Holmes, 2001). Micceri (1989) found that most psychological data samples do not meet the assumptions of parametric analyses. Others have found that Micceri's general caveat applies to many neuro-imaging data sets (Faux & McCauley, 1990; Karniski et al., 1994). When used with real PET data, MPTs are at least as powerful as comparable parametric analyses (Arndt et al., 1996). Additionally, the power of MPTs may be improved over that shown in the present study by using pooled variance estimates to compute MPT test statistics (Holmes et al., 1996) and by using substantially more permutations (e.g., 10,000) than were used in the present study (Jockel, 1986). In summary, multivariate permutation tests offer a method of controlling for multiple testing while maximizing statistical power. This approach is particularly important in fields that regularly use small sample sizes, multiple significance testing to address the same research question, and data sets that frequently violate the assumptions of parametric analyses. Electrocortical studies were offered as an example of such a field, but many other areas of study have these qualities and would also benefit from the use of MPTs.

ACKNOWLEDGMENTS

The work for this article was conducted while the authors were partially supported by the NICHD core grant HD15052 (to John F. Kennedy Center, Vanderbilt University).

REFERENCES

Arndt, S., Cizadlo, T., Andreasen, N., Heckel, D., Gold, S., & O'Leary, K.D. (1996). Tests for comparing images based on randomization and permutation methods. *Journal of Cerebral Blood Flow and Metabolism*, *16*, 1271–1279.

Blair, R., Higgins, J., Karniski, W., & Kromrey, J. (1994). A study of multivariate permutation tests which may replace Hotelling's T^2 test in prescribed circumstances. *Multivariate Behavioral Research*, *29*, 141–159.

Blair, R.C., & Karniski, W. (1994). Distribution-free statistical analyses of surface and volumetric maps. In R. Thatcher, M. Hallet, T. Zeffiro, E. R. John, & M. Huerta (Eds.), *Functional neuroimaging: Technical foundations* (pp. 19–28). San Diego: Academic Press.

Bonferroni, C.E. (1950). Sulle medie multiple di potenze. *Bottettino dell'Unione Matematica Italiana*, *5*, 267–270.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249–253.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, M. (1987). Analysis of variance with repeated measures on autonomic responses. *Psychophysiology*, *24*, 475–476.

Dawson, G., Finley, C., Phillips, S., & Lewy, A. (1989). A comparison of hemispheric asymmetries in speech-related brain potentials of autistic and dysphasic children. *Brain and Language*, *37*, 26–41.

Dunn, O.J. (1959). Confidence intervals for the means of dependent, normally distributed variables. *Journal of the American Statistical Association*, *54*, 613–621.

Edgington, E.S. (1969). Approximate randomization tests. *Journal of Psychology*, *72*, 143–149.

Edgington, E.S. (1987). *Randomization tests* (2nd ed.). New York: Marcel Dekker.

Faux, S., & McCauley, R. (1990). Analysis of scalp voltage asymmetries using Hotelling's T^2 methodology. *Brain Topography*, *2*, 237–245.

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

Greenhouse, S., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*, 95–112.

Greenwald, A. (1993). Consequences of prejudice against the null hypothesis. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences* (pp. 419–448). Mahwah, NJ: Lawrence Erlbaum Associates.

Hayes, A. (1996). Permutation test is not distribution-free: Testing H_0: $p = 0$. *Psychological Methods*, *1*, 184–198.

Henderson, L., Yoder, P., Yale, M., & McDuffie, A. (2002). Getting the point: Electrophysiological correlates of proto-declarative point. *International Journal for Developmental Neurosciences*, *20*, 449–458.

Hochberg, Y., & Tamhane, A. (1987). *Multiple comparison procedures.* New York: Wiley & Sons.

Holmes, A., Blair, R., Watson, J., & Ford, I. (1996). Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow and Metabolism*, *16*, 7–22.

Huynh, H., & Feldt, L. (1970). Conditions under which mean square ratios in repeated measures design shave exact *F*-distributions. *Journal of the American Statistical Association*, *65*, 1582–1589.

Jockel, K. (1986). Finite sample properties and asymptotic efficiency of Monte Carlo tests. *Annals of Statistics*, *14*, 336–347.

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

Kramer, A.F., Trejo, L.J., & Humphrey, D. (1995). Assessment of mental workload with task-relevant auditory probes. *Biological Psychology*, *40*, 83–100.

Kromrey, J.D., & La Rocca, M.A. (1995). Power and Type I error rates of new pairwise multiple comparison procedures under heterogeneous variances. *The Journal of Experimental Education*, *63*, 343–362.

McCall, R., & Appelbaum, M. (1973). Bias in the analysis of repeated measures designs: Some alternative approaches. *Child Development*, *44*, 401–415.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166.

Miller, R. (1981). *Simultaneous statistical inference* (2nd ed.). New York: Springer.

Nichols, T., & Holmes, A. (2001). Nonparametric permutation tests for functional neuro-imaging: A primer with examples. *Human Brain Mapping*, *15*, 1–125.

Pesarin, F. (2001). *Multivariate permutation tests.* New York: Wiley & Sons.

Potts, G.F., Dien, J., Hartry-Speiser, A., McDougal, L., & Tucker, D. (1998). Dense sensor array topography of the event-related potential to task-relevant auditory stimuli. *Electroencephalography and Clinical Neurophysiology*, *106*, 444–456.

Tabachnick, B., & Fidell, L. (1996). *Using multivariate statistics* (3rd ed.). Northridge, CA: HarperCollins.

Trainor, L., Samuel, S., Desjardins, R., & Sonnadara, R. (2001). Measuring temporal resolution in infants using mismatch negativity. *Neuroreport*, *12*, 2443–2448.

van Laar, M.W., Volkerts, E.R., Verbaten, M.N., Trooster, S., Kenemans, J., & van Megan, H. (2002). Differential effects of amitriptyline, nefazondone, and paraoxetine on performance and brain indices of visual selective attention and working memory. *Psychopharmacology*, *162*, 351–363.

Vasey, M.W., & Thayer, J.F. (1987). The continuing problem of false positives in repeated measures ANOVA in psychophysiology: A multivariate solution. *Psychophysiology*, *24*, 479–486.

Westfall, P., & Young, S. (1993). *Resampling-based multiple testing*. New York: Wiley & Sons.