METHODOLOGY

An alternative method for significance testing of waveform difference potentials

R. CLIFFORD BLAIR* AND WALT KARNISKI*:

*Department of Community and Family Health, College of Public Health,
University of South Florida, Tampa
*Department of Pediatrics, College of Medicine, University of South Florida, Tampa

Abstract

Guthrie and Buchwald (1991) proposed an ad hoc procedure for assessing the statistical significance of waveform difference potentials that may arise in a variety of psychophysiology research contexts. In our paper, an alternative method is presented and demonstrated that has fewer underlying assumptions than does the Guthrie-Buchwald test and may, therefore, produce better results in some situations. In particular, the test proposed here (a) is distribution free, (b) requires no assumption of an underlying correlation structure (e.g., first-order autoregressive), (c) requires no estimate of the population autocorrelation coefficient, (d) is exact, (e) produces p values for any number of subjects and time points, and (f) is highly intuitive as well as theoretically justifiable. This procedure may be used to carry out multiple comparisons with exact specification of experimentwise error, however, this test is based on permutation principles and may require large amounts of computer time for its implementation.

Descriptors: Waveform analysis, Permutation tests, Significance testing

Most psychophysiological studies of evoked potentials assess the effect of a treatment condition on the evoked potential in one of two ways. Often a relatively small number of peaks in the difference waveform are identified by visual inspection, and the amplitudes of these peaks are subjected to statistical analysis. This method may introduce bias into the study by narrowing attention to specific peaks to the exclusion of other less impressive phenomena. Also, the point of maximum difference may not necessarily correspond to the point of greatest effect. Other methods rely on data reduction techniques, such as principal components analysis, to produce data points for further statistical study. These methods also require subjective judgments (e.g., the number of components to be extracted) that affect the selection or exclusion of data for analysis. Although the utility of such methods cannot be denied, it is prudent in some circumstances to analyze all waveform data points without prior selective screening.

Recently, Guthrie and Buchwald (1991) proposed a method that utilizes an entire waveform for evaluating the statistical significance of difference potentials obtained from a group of subjects. Such a procedure has obvious implications for the analysis of data collected in connection with studies of event-related brain potentials (ERPs) as well as for other research areas of interest to psychophysicists. For example, Guthrie and Buchwald consider a study of P3, which is produced in a frequent–rare paradigm with the difference potential attained as the difference between the average responses to frequent and rare conditions. In such a study, the researcher would ordinarily be interested in determining whether or not a statistically significant difference exists between the outcomes of the two treatment conditions.

The purpose of the present paper is to outline an alternative method for carrying out such a test that circumvents some of the potential difficulties associated with the Guthrie-Buchwald procedure. We briefly describe the Guthrie-Buchwald (G-B) procedure, propose an alternative testing method, demonstrate its application, and discuss some of the characteristics of the alternative test.

The G-B Test

Consider an experiment in which voltages are digitally recorded, following some stimulus, across a finite period of time (T), and such recordings are made for each of N subjects under each of two conditions. (These recordings are usually the average of several simple trials.) Differences between recordings for corresponding time points made under the two treatment conditions can then be used to evaluate treatment effect differences. If only one time point were involved, this evaluation could be made by means of an ordinary paired-samples t test. In the case of multiple time points, a paired-samples t test could be computed at
each such point, but a question then arises as to how a determination of statistical significance should be made. Because studies of this sort often involve a great number (100 or more) of such points, the insensitivity that would accrue to a strategy based on Bonferroni adjustments makes this an unattractive option.

Because differences at temporally proximate points are likely to be correlated, Guthrie and Buchwald reasoned that an effect such as P3 is likely to produce a series of adjacent \( t \) values that all fall into some prescribed critical region such as .05 or .10. For this reason, these authors suggested that the number of \( t \) statistics making up the longest observed series of this type be used as the test statistic. To generate critical values for their test statistic, Guthrie and Buchwald assumed a first-order autoregressive structure for the difference potentials and used this assumption to generate critical values by computer simulations. The resulting table of critical values was formulated for situations where \( T = 10, 25, 50, 100, \) or 150, \( N = 10, 15, 20, 25, \) or 30, and the autocorrelation is .5, .7, or .9. Computer-simulated estimates of critical values were based on 1,000 repetitions of each combination of conditions.

Although this table of critical values was constructed with a known value for the population autocorrelation coefficient, this value will not typically be available to the researcher, thereby necessitating its estimation from the sample data. When the number of time points or subjects used in a study does not match those provided in the table, some method of interpolation or extrapolation is necessary.

A Proposed Test

The theoretical justification for the tests proposed in this section is rooted in permutation test theory. Because some knowledge of this method of test construction is necessary to follow the discussion, we begin by outlining the rationale underlying the permutation version of the familiar univariate paired-samples \( t \) test. An example is given, and a simple test of significance is carried out. This example is then extended to the multivariate domain, and the proposed testing method is explained in that context.

Suppose that each of three subjects is tested under Conditions A and B (e.g., pre-post, rare-frequent, etc.). We assume a true null hypothesis for the moment, that is, no difference between conditions. The measures thus obtained are shown in Table 1 for Permutation 1. The \( t \) value of 3.46 is the paired-samples \( t \) statistic computed for these data.

If the null hypothesis is true, that is, if there truly is no difference between conditions, then the order of observations for a given subject is arbitrary. Thus, the third subject who earned scores of 6 and 4, respectively, under Conditions A and B is just as likely to have scores of 4 under A and 6 under B (because there is no difference between conditions). Because this argument can be made for each of the three subjects, there are 2\(^3\) = 8 different possible arrangements of the data. Each of these arrangements or permutations is equally likely if the null hypothesis is true. Each of these arrangements along with its associated \( t \) statistic is shown in Table 1. The key concept here is that, given the scores earned by the three subjects and a true null hypothesis, the eight data arrangements represent all possible outcomes of the experiment and hence all possible \( t \) values. This simple idea forms the logical rationale for permutation tests. We now relax our assumption of a true null hypothesis and conduct the significance test by first arranging the eight \( t \) values into ascending order as \(-3.46\), \(-1.11\), \(-0.46\), \(0.0\), \(0.0\), \(0.46\), \(1.11\), and \(3.46\). This list represents the reference distribution for the test statistic. Because the \( t \) value actually produced by the experiment was 3.46, we can now pose the (one-tailed) question "What is the probability of obtaining a \( t \) value of 3.46 or greater from this experiment if the null hypothesis is true?" A look at the reference distribution provides the answer of \(\frac{1}{8}\) or .125. The corresponding two-tailed \( p \) value is then \(\frac{2}{8}\) or .25. Thus, a one-tailed test of the null hypothesis would be rejected at the .125 level, and a two-tailed test would be rejected at the .25 level. (Because of symmetry of the reference distribution, these tests could have been conducted by constructing either half of the distribution.)

Critical values obtained in this manner are not estimates of the appropriate critical values, they are the actual appropriate critical values. Therefore, this test is an exact test. Also, at no point in the reasoning process outlined above was an appeal made to an assumption of population normality or any other population shape. Thus, the test is distribution free in the same sense that Wilcoxon's signed ranks test is distribution free. Edgington (1987) maintained that, unlike parametric statistics, permutation tests do not require an assumption that observations are randomly sampled from a population. However, this test is not without assumptions. It shares with its parametric counterpart the assumption that observations across subjects are independent of one another. A violation of this assumption would occur if the score earned by one subject were to influence or in some fashion depend upon the score earned by another subject.

Readers unfamiliar with permutation (or randomization) tests may harbor doubts about the validity of the logic outlined

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>8</td>
<td>5</td>
<td>8</td>
<td>5</td>
<td>8</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.46</td>
<td>0.46</td>
<td>1.11</td>
<td>0.0</td>
<td>0.0</td>
<td>-1.11</td>
<td>-0.46</td>
<td>-3.46</td>
</tr>
</tbody>
</table>

above and by extension to the test it produces. The question might well arise as to whether or not such tests can be expected to produce results similar to those obtained from more familiar parametric counterpart tests. However, this question must be reversed. As Bradley (1968, p. 85) pointed out, "eminent statisticians have stated that the randomization test is the truly correct one and that the corresponding parametric test is valid only to the extent that it results in the same statistical decision." Edgington (1987) included Sir Ronald Fisher and Oscar Kempthorne in this eminent group.

We now consider a multivariate extension of the above example. Suppose that instead of obtaining only one measure for each subject under Conditions A and B, we obtain two measures. In the context of the subject at hand, this would mean observing two time points rather than one. Table 2 shows these data. Again, we assume that the observations actually generated by the experiment are those shown under the Permutation 1 heading. Because two variables are now involved, it is possible to compute a paired-samples t test for each. These two statistics are shown in parentheses. The question now arises as to how these two quantities may be used to generate multivariate statistics that have useful properties. Let us first consider summation of the two t values. In the event that a treatment were to cause an increase (or decrease) in both of the variables being studied, then we would expect this change to be reflected in both t values and would further expect that both values would have the same algebraic sign. By summing these two statistics we would, in a sense, combine the information from each univariate result into a single multivariate statistic. We designate this statistic as t_{sum} in the table. If, however, the treatment were to bring about an increase in one variable and a decrease in the other, t_{sum} would not be an efficient or even appropriate statistic because summation of the differently signed univariate statistics would tend to cancel the effect of each. In this circumstance, we might consider summation of the absolute values of the two tests. We designate this statistic as |t_{sum}|. In contrast to the first two situations, the treatment may be likely to bring about a change in only one of the two variables, although we cannot designate which one. In this case we might use the t statistic with greatest absolute value for test purposes. We designate this statistic as t_{max}. Each of these statistics has been tailored for sensitivity to a specific form of treatment effect. (See Edgington [1987, chap. 7] for an additional discussion of t_{sum} and |t_{sum}|.)

A significance test using any one of the above multivariate statistics can be carried out in the same manner as that described for the univariate case. The reference distribution would be formed by placing the statistics obtained from each permutation in rank order. The statistic actually obtained in the experiment would then be referenced to this distribution, and a p value would be obtained.

In the multivariate case, the set of observations must be permuted for each subject, not the observations taken on the individual variables. This means, for example, that the values of 6 and 7 earned by the third subject under the A condition must be permuted together and never individually because observations taken on a given subject are likely to be correlated in some manner and so cannot be treated as independent. However, we need not take this relationship into account in any other manner to obtain a valid test. We need not, for example, estimate the correlation between these variables or make any form of assumption concerning the manner in which they are related.

The flexibility of this testing method is important. We can include the series type statistic proposed by Guthrie and Buchwald into this testing procedure. Although it is difficult to demonstrate with this small example, given a large number of observations on each subject we would simply calculate the series statistic on each permutation and proceed as usual with the testing procedure. Although we would be using the test statistic described by Guthrie and Buchwald, the significance test itself would have very different characteristics from those proposed by those authors. All of the statistics described in this section, including the series statistic, have the following characteristics.

The tests are exact in that they do not use estimates of critical values but rather employ the critical value itself. Also, there is no assumption concerning the underlying correlational structure of the data. This means that the pattern of correlations between observations may be strong, weak, regular, irregular, or of any other form—it makes no difference in the testing procedure and does not influence the validity of the test. Because no particular structure is assumed, there are no doubts concerning the effects on the validity of the test of violating some assumed structure. Therefore, the researcher need not be concerned with estimating the autocorrelation with the accompanying hope that the obtained estimate is accurate. All of these tests are also distribution free so that we need not assume any given population shape.

### Some Example Analyses

Extension of the two-time-points, three-subjects example to the case of arbitrary numbers of time points and subjects is conceptually, if not computationally, quite straightforward. Suppose, for example, that 200 time-point observations are taken on 15 subjects under each of two conditions. In this case, the

### Table 2. Eight Data Set Permutations for Multivariate Permutation Tests

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8.6</td>
<td>5.2</td>
<td>8.6</td>
<td>5.2</td>
<td>8.6</td>
<td>5.2</td>
<td>8.6</td>
<td>5.2</td>
</tr>
<tr>
<td>B</td>
<td>4.3</td>
<td>3.1</td>
<td>4.3</td>
<td>3.1</td>
<td>4.3</td>
<td>3.1</td>
<td>4.3</td>
<td>3.1</td>
</tr>
<tr>
<td>t</td>
<td>(3.46, 5.20)</td>
<td>(0.46, 0.48)</td>
<td>(1.11, 0.90)</td>
<td>(0.0, -0.15)</td>
<td>(0.0, 0.15)</td>
<td>(-1.11, -0.90)</td>
<td>(-0.46, -0.48)</td>
<td>(-3.46, -5.20)</td>
</tr>
<tr>
<td>t_{sum}</td>
<td>8.66</td>
<td>0.94</td>
<td>2.01</td>
<td>-0.15</td>
<td>0.15</td>
<td>-2.01</td>
<td>-0.94</td>
<td>-8.66</td>
</tr>
<tr>
<td></td>
<td>8.66</td>
<td>0.94</td>
<td>2.01</td>
<td>-0.15</td>
<td>0.15</td>
<td>-2.01</td>
<td>-0.94</td>
<td>-8.66</td>
</tr>
<tr>
<td>t_{max}</td>
<td>5.20</td>
<td>0.48</td>
<td>1.11</td>
<td>-0.15</td>
<td>0.15</td>
<td>-1.11</td>
<td>-0.48</td>
<td>-5.20</td>
</tr>
</tbody>
</table>
reference distribution would be formed by permuting the sets of 200 observations in the same manner as was done with the two observations in the simpler example. Two hundred paired-samples \( t \) statistics would be computed for each permutation, and the desired multivariate statistic (e.g., \( t_{\text{sum}} \), \( t_{\text{sum}}^2 \), \( t_{\text{max}} \), or the series statistic of Guthrie-Buchwald) would be calculated. These multivariate statistics would then be ordered and formed into the appropriate reference distribution as before. However, in general the number of permutations required for a specific application will be \( 2^n \) or, in this example, \( 2^{15} = 32,768 \). Even for this modest number of subjects, the power of a computer will be required to carry out the necessary calculations.

We will now conduct significance tests using the four multivariate test statistics on data collected in connection with actual event-related studies. The top panel of Figure 1 depicts averaged frequent and rare waveforms obtained from 13 subjects during the course of a study of the temporal stability of auditory P3 derived from a standard odd-ball paradigm using pure tones as stimuli. Each of these waveforms is made up of 500 time-point observations. The middle panel of Figure 1 shows the average difference potential obtained by subtracting frequent from rare waveforms. The task at hand is to determine whether or not there is a statistically significant difference between the two waveforms shown in the upper panel or, equivalently, to determine whether or not any of the points shown in the middle panel differ significantly from zero. To arrive at a determination, we apply the permutation tests.

Table 3 shows results obtained from application of the four permutation tests to the waveform data shown in Figure 1. This table shows the test conducted, the value of the test statistic obtained from the data, and the \( p \) values associated with the obtained test statistic for each comparison. Although \( t_{\text{sum}} \) produced a nonsignificant result, the other three tests were significant, indicating that frequent and rare ERPs differ in some respect. The nonsignificant result for \( t_{\text{sum}} \) would be expected because the P3 paradigm generates both positive and negative components. Summing over these differently signed components would tend to cancel the effect of each. Thus, a researcher would ordinarily not choose \( t_{\text{sum}} \) for testing in this context. For testing a hypothesis, the researcher must designate in advance the test to be used if test level is to be maintained. Ordinarily, a researcher should not observe the four results and then choose the one that best confirms a desired result. (Alternatively, all four tests could be considered if a Bonferroni-type adjustment were made to the associated \( p \) values.) However, this caution does not rule out use of the four tests for exploratory purposes. Comparison of test results can lend some insight into the nature of the obtained effect. For example, an effect of short duration might produce a significant result for \( t_{\text{max}} \) and nonsignificance for the other tests.

An interesting and important characteristic of \( t_{\text{max}} \) is demonstrated in the bottom panel of Figure 1. Suppose that, rather than being satisfied with the general test of hypothesis concerning no difference between waveforms, we wish to conduct a test

---

**Table 3. Results of Permutation Significance Tests on Waveform Comparisons**

<table>
<thead>
<tr>
<th>ERP comparison</th>
<th>Statistic</th>
<th>Value</th>
<th>( p ) (exact)</th>
<th>( p ) (approx.)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequent vs. rare</td>
<td>( t_{\text{sum}} )</td>
<td>-43.220</td>
<td>.412</td>
<td>.414</td>
</tr>
<tr>
<td></td>
<td>( t_{\text{sum}}^2 )</td>
<td>1,824.373</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>( t_{\text{max}} )</td>
<td>13.379</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>series</td>
<td>142.000</td>
<td>.002</td>
<td>.003</td>
</tr>
<tr>
<td>Frequent vs. frequent</td>
<td>( t_{\text{sum}} )</td>
<td>71.005</td>
<td>.387</td>
<td>.393</td>
</tr>
<tr>
<td></td>
<td>( t_{\text{sum}}^2 )</td>
<td>449.348</td>
<td>.351</td>
<td>.346</td>
</tr>
<tr>
<td></td>
<td>( t_{\text{max}} )</td>
<td>3.363</td>
<td>.170</td>
<td>.167</td>
</tr>
<tr>
<td></td>
<td>series</td>
<td>15.000</td>
<td>.528</td>
<td>.520</td>
</tr>
</tbody>
</table>

*Based on 10,000 random permutations.

---

**Figure 1.** Averaged frequent and rare waveforms obtained from 13 subjects in a study of P3 (top); average difference potential waveform obtained by subtracting frequent from rare waveforms (middle); plot of paired-samples \( t \) statistics computed at each time point (bottom).
at each of the 500 points to determine where differences occur. Conducting 500 individual t tests without proper adjustment would undoubtedly lead to an inflated experimentwise error (EWE) rate (the EWE rate is the probability of rejecting one or more of the individual tests under a true null hypothesis). In this circumstance, the reference distribution associated with \( t_{\text{max}} \) provides an efficient means of establishing and maintaining a desired EWE rate. By definition, each of the individual t statistics must be less than or equal to \( t_{\text{max}} \) in absolute value. Rejection of the null hypothesis for one or more of the individual tests at a prescribed level occurs only when \( t_{\text{max}} \) is rejected at that same level, which implies that the probability of rejecting one or more of the individual tests is equal to the probability of rejecting \( t_{\text{max}} \). Thus, if we wish to conduct two-tailed tests while maintaining EWE at .05, we need only find the value in the \( t_{\text{max}} \) reference distribution that cuts off .025 in each tail and compare each of the 500 t statistics with this value to make a determination of significance. In this example, the required critical value is \( \pm 4.686 \). The bottom panel of Figure 1 shows a plot of the 500 individual t statistics along with horizontal lines depicting the two-tailed critical values. Points falling above 4.686 or below \(-4.686\) can be declared statistically significant with assurance that experimentwise error will be maintained at .05.

This method of controlling Type I error inflation can be compared with that afforded by Bonferroni adjustments. For the problem at hand, the Bonferroni method would be implemented by comparing each test with the critical value which cuts off .025/500 = .00005 in each tail of a t distribution with \( N - 1 = 12 \) degrees of freedom. This critical value is 5.694, which is considerably larger than the 4.686 used for the permutation test. The Bonferroni method does not maintain EWE at a prescribed level but only guarantees that EWE will not exceed the prescribed level. Because actual EWE may be considerably smaller than that desired by the researcher, the Bonferroni method can be unduly conservative, resulting in significant power losses. This circumstance is particularly likely when data points are highly correlated, as is usually true of waveform data. In contrast, the permutation method results in an exact test, which is no more conservative than is necessary to maintain EWE at the desired level.

In contrast to the data presented in Figure 1, the top and middle panels of Figure 2 show data collected from the same 13 subjects, where both waveforms represent frequent conditions obtained 15 min apart. Because both waveforms were collected under the same treatment condition, no difference is expected between the two. Table 3 shows that all tests produce nonsignificant results. Likewise, Figure 2 shows that none of the 500 individual t values approach the \( t_{\text{max}} \) critical level. The critical levels depicted in Figure 2 differ from those in Figure 1 because the critical value, being exact, is adjusted to reflect the correlation structure of the data. The Bonferroni critical value, which is insensitive to the data structure, remains unchanged.

Figure 2. Averaged frequent waveforms obtained 15 min apart from 13 subjects (top); average difference potential waveform obtained by subtracting first from second frequent waveforms (middle); plot of paired-samples t statistics computed at each time point (bottom).

Concluding Remarks

The strengths and weaknesses of the permutation procedures discussed in this paper can be compared with those of the G-B test. The permutation tests do not require the burdensome assumption of a particular underlying correlational structure to assure maintenance of a desired Type I error rate, which is important because it is unlikely that any single reference model will be appropriate for the varied situations involved in ERP research. The paucity of information concerning the robustness (or lack thereof) of the G-B test to departures from model assumptions highlights the importance of this feature.

Relaxation of the assumption of a particular underlying correlational structure also precludes the necessity of estimating the
unknown autocorrelation parameter. This is important for at least two reasons. First, the estimation of this parameter may introduce an unknown amount of error into the testing procedure, even if the estimator were known to have good statistical properties (e.g., unbiased, small mean square error), which has not been demonstrated. Second, estimation of this parameter can be very complex, as outlined by Guthrie and Buchwald (1991, appendix). Although psychophysicists with significant relevant statistical training will have little difficulty implementing this estimation procedure, others may not find it an easy task.

Inspection of Guthrie and Buchwald's appendix also reveals that their test has an underlying assumption of difference potential population normality at each time point. Because difference potentials are likely to be symmetric under a true null hypothesis, a condition conducive to robustness for the paired-samples t test (Sawilowsky & Blair, 1992), we assume, without other evidence, that violation of this assumption will not significantly affect this test. However, the permutation tests do not require this assumption.

Use of estimated, rather than exact, critical values introduces an unevaluated amount of error into the G-B testing procedure. All permutation tests utilize exact critical values, thereby eliminating this source of error. There are circumstances, however, where it may be necessary to use estimated critical values in conjunction with the permutation tests. However, with these tests this source of error can be controlled and thereby maintained at a low level.

Critical values for permutation tests are explicit insofar as number of time points and subjects are concerned. Thus, researchers using these methods need not be concerned regarding a mismatch between numbers of time points and subjects in an experiment and values for these factors offered in a table of critical values, thus eliminating another potential source of error from consideration. A related benefit to be derived from this method of significance testing rests in the flexibility to be gained in choosing significance levels. As presently constituted, the G-B table forces tests to be conducted at the .05 level. Although tests at this level are most common, circumstances often require that other levels be employed. Not only may researchers sometimes wish to maintain more stringent control of Type I errors, but power considerations may also cause a researcher to conduct more liberal tests in some circumstances. Because the permutation procedures produce a p value rather than simply comparing an obtained statistic with a critical value, researchers using these tests are free to choose the test level that is most appropriate for their research setting.

The relative merits of the permutation tests discussed here can be compared with those of Hotelling's one-sample $T^2$ test as they relate to the analytic problem at hand. The $T^2$ test cannot be directly applied to the 500-point waveform array because computation of this statistic cannot be carried out when the number of data points exceeds the number of subjects. Even if a significant reduction in data were realized through the averaging of adjacent data blocks or some other method, this test may still not be appropriate because it is often very power when the number of subjects is only slightly greater than the number of variables. (See Blair, Higgins, Karnisky, & Kronrey, 1992, for power comparisons.) Unlike the permutation tests, the $T^2$ is also burdened by its associated assumption of population multivariate normality. A potential advantage of Hotelling's test is that it is designed to detect general departures from the null condition, whereas the permutation tests are tailored to more specific forms of the alternative. The $T^2$ test may detect treatment effects that were unanticipated by the researcher and are of forms that are not easily detected by the permutation tests. The reverse is also true, however, because the $T^2$ test can be quite insensitive to certain common treatment effect forms (O'Brien, 1984). The $T^2$ test is only marginally useful for the many-data points, few-subjects type of problem considered here.

Critical values associated with $f_{max}$ can be used to perform multiple comparisons with exact specification of EWE rates. Although the example presented involved only 13 subjects and 500 time points, reasonable sensitivity was attained. More sensitive tests will be obtained in situations that involve more subjects and fewer time points. Further studies of the power properties of this method for performing multiple comparisons are planned, and this method can be extended to other testing situations where correlated data are involved.

A more subtle but quite important positive characteristic of these tests is the logic underlying permutation testing procedures can be fully comprehended without appeal to mathematical sources that may be unavailable to researchers wishing to apply such tests. Unlike other forms of significance testing, the researcher is not required to apply a "black box" solution to the testing problem with little understanding of the intricacies or ramifications of the solution thus obtained. An understanding of the test greatly increases the probability that it will be appropriately and judiciously applied and means that researchers can take an active role in the development of test statistics that may be more sensitive to alternatives encountered in their particular area of research.

The very important advantages of permutation tests are purchased at significant, though manageable, costs. These costs are computationally related. Historically, the sheer volume of computations required for analytic problems of even moderate size made for an impossible task before the development of computers and was often prohibitively expensive even after their development. In recent years, two developments have greatly ameliorated this problem. Computers have now reached a level of sophistication and power to permit permutation problems of significant size to be solved on commonly available desktop models. For example, the computations carried out to produce the exact $p$ values reported in Table 3 required about 28.9 min of computing time on a 33-mz. 8088686 desktop computer and about 2.7 min on a RISC-based desktop system. The amount of computer time required increases rapidly as the number of subjects increases, however, so that a limit must eventually be reached beyond which such tests are not feasible. However, improvements in computers make yesterday's limit today's very manageable problem. New desktop machines, such as the RISC-based computers now being offered by a number of companies, have dramatically increased the magnitude of permutation data analysis problems that may be conveniently addressed on desktop systems. The researcher who has access to mainframe "super computers" can manage much larger problems.

Even when available computing power is moderate, researchers can still use permutation methods to address large analysis problems through approximate randomization tests (Edgingtont, 1987). The logic of these tests is the same as that discussed for the exact tests, except that instead of forming all possible permutations, a large number of random permutations is formed. Fortunately, these approximate permutation tests produce results very similar to those generated by exact methods as long as the
number of random permutations is large enough.\footnote{An IMSL-based Fortran program that performs exact and approximate versions of the tests discussed in this paper is available from the first author. An SAS program that implements only the approximate versions is also available.} In our data analysis problems, we typically compute both exact and approximate tests to study the characteristics of the approximate tests. A set of 10,000 random permutations produces very satisfactory results for tests of the sort discussed in this paper. In Table 3, the approximate $p$ values agree quite closely with the exact values. Large scale Monte Carlo studies of the Type I error and power properties of the approximate test versions of $t_{\text{sum}}$, $t_{\text{sum}}$, and $t_{\text{max}}$ have been conducted and reported by Blair et al. (1992). Results were quite good on both counts.

Researchers in psychophysiology often face statistical problems for which existing solutions are inadequate or for which no reasonable solution exists at all. In this paper, we have followed the lead of Guthrie and Buchwald (1991) in attempting to provide inroads into one such problematic area.

**REFERENCES**


(Received July 9, 1991; Accepted June 25, 1992)
This document is a scanned copy of a printed document. No warranty is given about the accuracy of the copy. Users should refer to the original published version of the material.