DESIGN AND ANALYSIS A Researcher's Handbook

Second Edition

GEOFFREY KEPPEL

Department of Psychology
University of California, Berkeley

Prentice-Hall, Inc., *Englewood Cliffs, New Jersey 07632*

9003

Library of Congress Cataloging in Publication Data

KEPPEL, GEOFFREY.
Design and analysis.

Bibliography: p.
Includes index.
1. Social sciences—Statistical methods. 2. Factorial experiment designs. 3. Social sciences—Research.
I. Title.
HA29.K44 1982 300'.72 81-15419
ISBN 0-13-200048-2 AACR2

Editorial/production supervision by
Alison D. Gnerre and Dee Amir Josephson
Cover design by Martha Geering
Manufacturing buyer: Edmund W. Leone

© 1982, 1973 by Prentice-Hall, Inc.,
Englewood Cliffs, New Jersey 07632

All rights reserved. No part of this book
may be reproduced in any form or by
any means without permission in writing
from the publisher.

Printed in the United States of America

10 9 8 7 6

Prentice-Hall International, Inc., *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall of Canada, Ltd., *Toronto*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Prentice-Hall of Southeast Asia Pte. Ltd., *Singapore*
Whitehall Books Limited, *Wellington, New Zealand*

8

Correction for Multiple Comparisons

8.1 Problems Associated with Analytical Comparisons (p. 145)
Increase in Type I Error
Planned Comparisons
Post Hoc Comparisons
8.2 Familywise Correction for Comparisons: The Scheffé Test (p. 151)
8.3 Comparisons Between Pairs of Means (p. 153)
The Dunnett Test
The Tukey Test
The Fisher Test
8.4 Practical Points to Consider (p. 159)
Conditionalized Planned Comparisons
A Decision to Suspend Judgment
Recommendations and Guidelines
8.5 Exercises (p. 166)

One of the most complex and perhaps most confused topics in statistical inference concerns an unfortunate by-product of the conscientious and detailed analysis of an experiment, namely, an increased vulnerability to type I error. That is, type I errors *cumulate* with each statistical test performed in a predictable and unavoidable fashion—unavoidable, unless steps are taken to correct this process. This chapter examines this problem and solutions designed to solve it. My coverage will be selective, focusing on particular recommendations rather than presenting a comprehensive description of the topic. If you wish to study the problem in more depth, you might read first two early discussions (Ryan, 1959; Wilson, 1962) and then three informative and useful papers by Games (1971b, 1978a, and 1978b).[1]

8.1 PROBLEMS ASSOCIATED WITH ANALYTICAL COMPARISONS

Increase in Type I Error

The major problem resulting from the performance of a series of analytical comparisons on a set of data is the unpleasant fact that the more comparisons we conduct, the more type I errors we will make when the null hypothesis is true. In talking about this relationship, the distinction is often made between the type I error per comparison (*PC*) and the error rate familywise (*FW*).[2] The *PC* error rate, which we will continue to call α, uses the *comparison* as the conceptual unit for the error rate. If we evaluated several comparisons in an experiment, each at $\alpha = .05$, we would be using a *PC* error rate; our probability of making a type I error would be .05 for each of the separate comparisons. In contrast, the type I *FW* error rate, considers the probability of making *one or more* type I errors in the *set of comparisons* under scrutiny.

The relationship between the two error rates is expressed by the formula

$$\alpha_{FW} = 1 - (1 - \alpha)^c, \tag{8-1}$$

where c represents the number of orthogonal comparisons that are conducted. With the error rate set at $\alpha = .05$ and with $c = 3$ comparisons contemplated, the *FW* error rate is

$$\alpha_{FW} = 1 - (1 - .05)^3 = 1 - (.95)^3 = 1 - .857 = .143.$$

Working at the 1 percent significance level,

$$\alpha_{FW} = 1 - (1 - .01)^3 = 1 - (.99)^3 = 1 - .970 = .030.$$

[1] Papers should be read in conjunction with Games (1971a), which corrects a number of typical errors in the original article.

[2] At the recommendation of others, I have adopted the term *familywise* to replace the term *experimentwise*. Both terms refer to the same concept in the context of the single-factor design. The distinction between the two terms becomes critical in factorial designs.

The familywise error rate is approximated by

$$\tilde{\alpha}_{FW} = c(\alpha), \tag{8-2}$$

but the approximation $\tilde{\alpha}_{FW}$ will always overestimate α_{FW}, though less for small values of α and for small numbers of comparisons. In the present example, with $\alpha = .05$,

$$\tilde{\alpha}_{FW} = 3(.05) = .15,$$

as compared with the $\alpha_{FW} = .143$ obtained with Eq. (8-1). For the smaller α level, $p = .01$, the two values are identical when carried to two decimal places.

When several comparisons are involved in the analysis of an experiment, researchers have different attitudes about which conceptual unit for the error rate is most appropriate. The point to be emphasized now is that when we conduct a number of comparisons on a set of means, there is an *FW* error rate with which we must contend and this error rate increases directly with the number of comparisons tested.

The relationship between *FW* error rate and the number of analytical comparisons is not an exclusive property of orthogonal comparisons, however, but holds for nonorthogonal comparisons as well. While the calculation of *FW* rate when nonorthogonal comparisons are involved in an analysis plan is not simple, it is still accurate to say that the *FW* error rate increases with the number of comparisons conducted regardless of orthogonality (see Harter, 1957, for a discussion of the problem).

Planned Comparisons

Experiments are usually designed with specific hypotheses in mind, and most researchers conduct analyses relevant to these hypotheses directly without reference to the outcome of the omnibus *F* test. Although the omnibus test may be computed, its significance or nonsignificance does not modify this particular course of action. Unplanned comparisons, in contrast, refer to the analysis of *unexpected* findings, outcomes that are revealed only after the data have been collected and partially analyzed. These post hoc comparisons, as they are also called, can sometimes be critically important in the development of a field of research.

While both planned and post hoc comparisons lead to an increase in *FW* error, they are usually treated differently in any attempt to reduce or to control the *FW* error rate. The most widely used strategy is to evaluate planned comparisons in the normal way—at the usual *PC*, or α, rate—and to exercise control of the *FW* rate for post hoc comparisons through special evaluation procedures designed to cope with the problem.[3]

[3]Not all commentators on the problem of *FW* error agree that planned comparisons should be left uncorrected, however. The correction they usually recommend for planned comparisons is the Bonferroni, or Dunn, test, which applies a less severe correction than those associated with tests developed for use with post hoc comparisons. The Bonferroni test is discussed in detail by Kirk (1968, pp. 79–81) and Myers (1979, pp. 298–300).

Planned comparisons are specified in the initial planning of an experiment. Collectively, they represent the answers an investigator hopes to obtain to the research questions that generated the experiment in the first place. Although the definition of planned comparisons is clear, the restrictions suggested by various authors are not. The two major areas of concern are the orthogonality of the set of planned comparisons and the number of comparisons to be included in the set. I will consider both issues in some detail.

The Issue of Orthogonality. I have already discussed this issue in Chapter 6 (pp. 121-123). Consensus seems to favor a set of meaningful comparisons which more often than not contains some nonorthogonal comparisons, i.e., a set of comparisons that are not mutually orthogonal. As you saw in Chapter 6, complete orthogonal sets do not necessarily contain comparisons that are all psychologically meaningful, on the one hand, nor do they completely enumerate the total set of meaningful comparisons associated with an experiment, on the other. The value of orthogonal comparisons lies in the independence of *inferences,* which, of course, is a desirable quality to achieve. That is, orthogonal comparisons are such that any decision concerning the null hypothesis representing one comparison is uninfluenced by the decision concerning the null hypothesis representing any other orthogonal comparison (see pp. 118-119).

The potential difficulty with nonorthogonal comparisons, then, is in interpreting the different outcomes. If we reject the null hypotheses for two nonorthogonal comparisons, which comparison represents the "true" reason for the observed differences? If we reject one null hypothesis and not the other, what can we make of the results? I do not mean to say that such ambiguities always arise when nonorthogonal comparisons are involved or that they are unresolvable when they do, but simply to suggest that nonorthogonal comparisons be interpreted with particular care in order to avoid this sort of problem.

The Number of Planned Comparisons. While everyone seems to agree that planned comparisons should be limited in number, there is no agreement on what the number should be. One obvious possibility is to restrict the number of comparisons to the number of degrees of freedom associated with the treatment source of variance ($df_A = a - 1$). But this suggestion, or any other recommendation for a smaller, is an *arbitrary* restriction. On the other hand, the point I wish to make is simply that many researchers do limit the number of planned comparisons and, depending on the research hypotheses and on the complexity of the experiment, the number of planned comparisons will range slightly above or below df_A. A conscientious researcher will realize that the number of comparisons affects the *FW* rate, however, and will reach a threshold of concern for this fact somewhere near the "natural" limit set by the number of degrees of freedom.

The Modified Bonferroni Test. One possible solution to this problem is to assume that most researchers are willing to test up to $a - 1$ planned comparisons without *correcting* for *FW* error. Suppose we use this number—the *df* associated with the between-groups mean square—to set the *FW* standard for planned comparisons. To introduce corrections only when the number of comparisons *exceeds*

df_A. This can be accomplished easily by calculating the α_{FW} associated with $a - 1$ orthogonal comparisons and then dividing this probability by the *actual* number of planned comparisons included in the analysis plan.[4] The resulting probability is a new *PC* rate to be used in assessing these comparisons that maintains α_{FW} at this presumably acceptable standard. In symbols, we use Eq. (8-2) to calculate the maximum *FW* error for planned comparisons,

$$\tilde{\alpha}_{FW\,\text{planned}} = (df_A)(\alpha), \tag{8-3}$$

and divide this value by the number of comparisons *actually planned* (c):

$$\tilde{\alpha}_{\text{planned}} = \frac{\tilde{\alpha}_{FW\,\text{planned}}}{c}. \tag{8-4}$$

This method of adjusting the *PC* rate is related to a procedure known as the **Bonferroni test**, which I mentioned in footnote 3 (p. 146). In the Bonferroni test, the numerator in Eq. (8-4) is the overall *FW* rate adopted by a researcher for the experiment—*not* the $\tilde{\alpha}_{FW\,\text{planned}}$ I have suggested as a method for controlling *FW* error with planned comparisons. I will refer to the present technique as the **modified Bonferroni test.**

As an example, consider an experiment with $a = 5$ conditions. Assuming that the maximum number of planned comparisons that will be conducted without correction is equal to $df_A = 4$, and adopting the standard significance level $\alpha = .05$, we obtain an approximate probability for this "acceptable" *FW* rate by substituting in Eq. (8-3). That is,

$$\tilde{\alpha}_{FW\,\text{planned}} = 4(.05) = .20.$$

This *FW* rate of .20, then, represents the assumed risk that researchers are generally willing to take when planned comparisons are at stake.

Suppose we wanted to conduct $c = 5$ planned comparisons. We find from Eq. (8-4) that

$$\tilde{\alpha}_{\text{planned}} = \frac{.20}{5} = .04$$

is the adjusted significance level that will be used to evaluate the significance of *all five* planned comparisons. A quick calculation will indicate that this new rejection probability results in an *FW* rate of $c(\tilde{\alpha}_{\text{planned}}) = 5(.04) = .20$, which is the *FW* rate we adopted in this example for the entire set of planned comparisons.

Table 8-1 summarizes the calculations of $\tilde{\alpha}_{\text{planned}}$ for several values of c using $a = 5$. You will note that when $c \leqslant df_A$, no correction is applied and the *FW* varies with the number of comparisons. On the other hand, when $c > df_A$, the *FW* rate remains constant at $\tilde{\alpha}_{FW\,\text{planned}} = .20$ and the significance level used to evaluate the comparisons is systematically lowered as a consequence.

[4]These calculations are based on orthogonal comparisons, which will tend to cause overcorrection of the *FW* error when nonorthogonal comparisons are involved, but this "overcorrection" makes little practical difference in an actual research application.

Table 8-1 The Modified Bonferroni Test for Planned Comparisons

NUMBER OF COMPARISONS (c)	REJECTION PROBABILITY ($\tilde{\alpha}_{\text{planned}}$)	APPROXIMATE FAMILYWISE ERROR ($\tilde{\alpha}_{FW}$)
1	.05	.05
2	.05	.10
3	.05	.15
4	.05	.20
5	.040	.20
6	.033	.20
7	.029	.20
8	.025	.20
9	.022	.20
10	.020	.20

One difficulty with the Bonferroni test is the need to determine the critical value of F when $\tilde{\alpha}_{\text{planned}}$ falls between the probabilitites provided in the standard F tables. Since the F's associated with most planned comparisons will involve 1 df in the numerator, it is possible to use the unit normal distribution and approximate the value of t, which, when squared, will give us the corresponding critical value of F. The value of t at an α level of significance is found by the formula

$$t(df_{S/A}) = z + \frac{z^3 + z}{4(df_{S/A} - 2)}, \tag{8-5}$$

where z represents the point on the unit normal distribution above which [$\frac{1}{2}(\alpha) \times 100$] percent of the curve falls.[5]

As an example, I will work with a calculation we can verify in the F table. From Table 8-1, you can see that $c = 8$ planned comparisons offers such a choice ($\tilde{\alpha}_{\text{planned}} = .025$). From a table of the unit normal curve, available in most introductory texts, the value of z above which $\frac{1}{2}(.025) \times 100 = 1.25$ percent of the area of the curve falls is 2.24. Substituting in Eq. (8-5) and assuming $df_{S/A} = 40$,

$$t(40) = 2.24 + \frac{(2.24)^3 + 2.24}{4(40 - 2)}$$

$$= 2.24 + \frac{11.24 + 2.24}{152} = 2.33.$$

For the goodness of the approximation, we square this value to give us

$$F(1, 40) = [t(40)]^2 = (2.33)^2 = 5.43.$$

The $F(1, 40)$ from Table A-1 in Appendix A at $\alpha = .025$ is 5.42.

[5] z provides a critical value of z that locates one half of the rejection region in the positive tail and the other one half of the rejection region in the negative tail of the unit normal

Summary. The decision to restrict the number of planned comparisons or to introduce some correction in the decision process depends on our attitude concerning type I and type II errors and the sort of balancing that we want to achieve between them. You must work this problem out for yourself and then deal with the additional problem of convincing others that your findings will hold up on replication. In short, part of your planning should include a concern for planned comparisons and your attitude toward them, as well as a concern for the logic of the experimental design, and a realistic estimate of power for the experiment in general and for the planned comparisons in particular. The implicit standard adopted by most researchers of conducting up to $a - 1$ planned comparisons without special correction seems reasonable except, perhaps, when the number of treatment conditions is particularly large and the *FW* error for planned comparisons becomes sizable as a consequence. Planned comparisons *are* special and should be evaluated with a sensitive statistical test. The modified Bonferroni test described in the preceding paragraphs offers a way of maintaining this standard in situations where the number of comparisons exceeds this assumed limit.

Post Hoc Comparisons

Post hoc comparisons often take the form of an intensive "milking" of a set of results—e.g., the comparison of all possible pairs of treatment means. The motivation, of course, is to extract the maximum amount of information from any given study. Another reason for conducting post hoc comparisons is that the results of such tests often lead to future experiments. An interesting comparison, significant or not, may form the basis for a new experiment. In the next study, for example, we might choose to manipulate more extensively the different treatments contributing to the comparison we have isolated.

The total number of possible single-*df* comparisons, which includes differences between pairs of means and complex comparisons between means, is staggering even for a "modest" experimental design. For $a = 3$, there are 6 such comparisons possible; for $a = 4$, there are 25; for $a = 5$, there are 90; and for $a = 6$, there are 301.[6] You can appreciate the concern for the *FW* rate when the pool of potential comparisons is as large as these!

There are several approaches designed to deal directly with this problem and such techniques employ the same basic solution, namely, to reduce the size of the critical region—i.e., to lower the significance level. In fact, Games (1971b) points out that the different procedures all involve the same underlying test statistic and differ only in the ways by which this reduction in the *PC* rate is achieved. The logic is straightforward: If we make it more difficult to reject the null hypothesis for each comparison tested, which must happen when the size of the rejection region is reduced, fewer type I errors will be committed and the *FW* error rate will thus be lowered. Just how much of an "adjustment" is made depends on a number

[6]The total number of comparisons is $1 + (3^a - 1)/2 - 2^a$.

of factors, such as our willingness to make type I errors in general, the number of post hoc comparisons actually conducted, and the pool of comparisons from which comparisons are specifically chosen. As you will see when you enter the research arena, there is no general agreement among researchers or even among authors of statistical texts and articles concerning these points.

In the next two sections, I will consider a number of procedures that have been developed to cope with the increased *FW* error rate associated with post hoc comparisons. Section 8.2 describes a technique that can be used to control the *FW* rate for any type of comparison between means. Section 8.3 focuses on techniques that have been designed specifically for comparisons between pairs of treatment means. I will not cover all of the tests that have been proposed or are even currently in use but will concentrate on those that for various reasons seem best suited to our needs. I have been assisted in this selection of tests by the analysis and recommendations of Games (1978b).[7]

8.2 FAMILYWISE CORRECTIONS FOR ALL COMPARISONS: THE SCHEFFÉ TEST

The Scheffé test is a technique that allows a researcher to maintain the *FW* rate at a particular value regardless of the number of comparisons actually conducted. For this reason, then, the Scheffé test is flexible in its application to the analysis of an experiment.

The Scheffé test requires no special tables, since it is based on the values of the *F* statistic appearing in standard *F* tables. The procedure is simple. We calculate $F_{\text{comp.}}$ in the usual fashion but evaluate the significance of the obtained *F* with a special critical value, F_S. This quantity is defined as follows:

$$F_S = (a - 1)F(df_A, df_{S/A}), \tag{8-6}$$

where $F(df_A, df_{S/A})$ is the critical value of *F* for the *omnibus* analysis of variance and is found in Table A-1 under the desired α level. (Be sure to note that the value of $df_{\text{num.}}$ is equal to $a - 1$ and *not* to $df_{\text{comp.}}$—a common mistake made by students in my classes.) The choice of significance level at this point sets the maximum value that familywise error α_{FW} may take regardless of the number of comparisons considered. Thus, an *F* chosen from Table A-1 at the .05 level of significance and substituted into Eq. (8-6) will create a cricital value of *F* (F_S) that sets a maximum limit on the *FW* rate at .05. Using critical values of *F* at other α levels sets the limit at other probabilities correspondingly.

As an example, suppose we performed an experiment with $a = 5$ treatment conditions and $s = 9$ subjects assigned randomly to each group. We will assume that $MS_{S/A}$, which is based on $a(s - 1) = 5(9 - 1) = 40$ degrees of freedom, is

[7] For an extensive summary and comparison of these and other techniques, see Kirk (1968, pp. 90–93) and Winer (1971, pp. 185–204).

13.22. The sums and the means for the treatment groups are presented in Table 8-2. Suppose we decided to compare the average of two of the groups (a_1 and a_5) with the average of the other three (a_2, a_3, and a_4). The first average is (12.78 + 11.44)/2 = 12.11, and the second average is (7.89 + 7.11 + 8.78)/3 = 7.93, indicating a sizable difference between the two means (12.11 − 7.93 = 4.18). A convenient set of coefficients with which to calculate the comparison sum of squares is (3, −2, −2, −2, 3). Substituting the necessary values in Eq. (6-5), we have

$$SS_{A\,\text{comp.}} = \frac{s(\hat{\psi})^2}{\Sigma\,(c_i)^2}$$

$$= \frac{9[(3)(12.78) + (-2)(7.89) + (-2)(7.11) + (-2)(8.78) + (3)(11.44)]^2}{(3)^2 + (-2)^2 + (-2)^2 + (-2)^2 + (3)^2}$$

$$= 189.00.$$

The next operation is to form an F ratio. Since we are still contrasting only two means in this comparison, the number of df for the $SS_{A\,\text{comp.}}$ is 1 and the $MS_{A\,\text{comp.}}$ = 189.00/1 = 189.00. The F ratio is specified in Eq. (6-7) and consists simply of dividing the $MS_{A\,\text{comp.}}$ by the $MS_{S/A}$. In this case,

$$F_{\text{comp.}} = \frac{189.00}{13.22} = 14.30.$$

Normally this F would be compared with the critical value of $F(1, 40)$, which at $\alpha = .05$ is 4.08. For the Scheffé test, however, we determine the critical value F_S by substituting in Eq. (8-6), for which we will need the critical value for the omnibus F, which in this case is $F(4, 40) = 2.61$ at $\alpha = .05$. Substituting in Eq. (8-6), we find

$$F_S = (a - 1)F(df_A, df_{S/A})$$
$$= (5 - 1)(2.61) = 10.44.$$

Since the obtained $F_{\text{comp.}}$ of 14.30 exceeds this critical value demanded by the Scheffé test (10.44), we can reject the null hypothesis.

Perhaps you noticed the severity of the Scheffé correction, which is reflected in the difference between the two critical values, 4.08 versus 10.44. Translated to a PC rate, the Scheffé correction is equivalent to a significance level of about α =

Table 8-2 Numerical Example: Treatment Sums and Means

	LEVELS				
	a_1	a_2	a_3	a_4	a_5
Sums	115	71	64	79	103
Means	12.78	7.89	7.11	8.78	11.44

.0025. This marked reduction in the *PC* rate (from $\alpha = .05$ to $\alpha = .0025$) is necessary to set the *FW* error at a value no greater than $\alpha_{FW} = .05$. When the frame of reference consists of all possible comparisons, however, which in this case is 90, the correction needs to be severe.

8.3 COMPARISONS BETWEEN PAIRS OF MEANS

There are times when a researcher may be interested in evaluating the significance of all possible differences between pairs of treatment means. The number of such pairwise comparisons, as they are frequently called, may be determined by solving the following simple formula: $a(a - 1)/2$. Applied research is often of this type—different books are compared in a classroom setting to determine which book is best; a consumer testing agency evaluates a number of similar products and attempts to order and to group the products in terms of effectiveness; and so on. In these cases, the intent is clearly to compare each treatment condition—each book or product, for example—with all other treatment conditions. Presumably there would be little or no interest in more complex comparisons unless there were good reasons for combining certain conditions, e.g., products from the same manufacturer or books by the same author. Theoretically motivated research, on the other hand, produces experimental designs that generate a limited number of meaningful comparisons that generally will *not* include all possible pairwise comparisons, but rather a smaller number of pairwise comparisons combined with a couple of more complex comparisons.

Many alternative tests have been developed to control *FW* error under the circumstances where a researcher wishes to conduct pairwise comparisons. The Scheffé test is *not* recommended for this situation, since it is less powerful than the more specialized techniques I consider in this section. I will cover three such tests in detail: the **Dunnett test**, which is used when the pairwise comparisons consist of differences between one condition (usually a control or baseline condition) and several others (usually experimental conditions); the **Tukey test**, which provides *FW* control when all of the pairwise comparisons are to be conducted; and the **Fisher test**, which offers *FW* control by means of an initial criterion that must be met before the comparisons can be conducted.

The Dunnett Test

When we include a control condition in an experiment, we are often interested in a number of different comparisons. As a first step in the analysis, we might contrast the control group with the average score for the experimental groups computing a sort of overall control-experimental contrast. Additionally, we might evaluate the significance of any differences observed among the experimental groups—a sort of omnibus *F* for the experimental groups. Finally, we would

probably consider multiple comparisons involving a contrast of each of the experimental groups with the single control group. Because of the necessary increase in the number of comparisons when a single control group is compared with several experimental groups, we might want to exercise some control over the *FW* error rate.

The Dunnett test is a specialized *FW* correction technique that compensates for the increased number of type I errors but is not as "corrective" as are other post hoc tests, because it takes into consideration only a limited number of comparisons—the control-experimental contrasts. The simplest way to conduct the Dunnett test is to calculate the control-experimental mean differences and to compare them against a critical mean difference ($\bar{d}_D$) that must be exceeded to be significant at the chosen *FW* level. The formula for calculating this critical difference is

$$\bar{d}_D = \frac{q_D \sqrt{2(MS_{S/A})}}{\sqrt{s}}, \tag{8-7}$$

where q_D refers to an entry in Table A-5 of Appendix A, $MS_{S/A}$ is the error term from the overall analysis of variance, and s is the sample size for each group. The value of q_D is determined by the total number of conditions k involved in the analysis, the degrees of freedom associated with the error term ($df_{S/A}$), and the value chosen for *FW* error (α_{FW}). If you choose to work with the F test, you can use

$$F_D = (q_D)^2 \tag{8-7a}$$

as the critical value with which to evaluate $F_{\text{comp.}}$.

As an example, I will use the data from Table 8-2. To set the example in context, let's assume that these data were drawn from an experiment comparing the amounts of memory loss for several different experimental conditions. There are four experimental groups (a_2, a_3, a_4, and a_5), each differing in the types of interfering activities they received between learning and recall. A control group (a_1) received a neutral task during the period in which the experimental subjects were experiencing interference.

While other questions might be asked of the data (e.g., questions about meaningful comparisons among the *experimental* groups), it is of interest to determine whether each of the experimental groups showed a significant loss relative to the control group. In order to calculate the critical C-E difference, we need to obtain q_D. Although the argument could be made that only differences in favor of the control group make any sense, most researchers would prefer to choose what is termed a nondirectional alternative hypothesis. What this means is that we want to be alert to positive as well as negative differences in the experiment. If we set the significance level at $q_{FW} = .05$, we will set aside half of the rejection region for positive deviations and the other half for negative deviations. Such a procedure is often called a **two-tailed test.**

To find the value of q_D, we locate the part of Table A-5 labeled "two-tailed

comparisons" and look for the entry at $k = 5$, $df_{error} = df_{S/A} = 40$, and $\alpha_{FW} = .05$. (A directional test would be conducted with the values given in the part of Table A-5 labeled "one-tailed comparisons.") For this combination, $q_D = 2.54$. Substituting in Eq. (8-7) gives as the critical difference between the control and an experimental mean

$$\bar{d}_D = \frac{(2.54)\sqrt{2(13.22)}}{\sqrt{9}} = 4.35.$$

This is the difference that must be exceeded in order to allow the rejection of the null hypothesis that the control group and a particular experimental group are equal. From Table 8-2 (p. 152), the observed differences are

$$\bar{A}_1 - \bar{A}_2 = 12.78 - 7.89 = 4.89; \qquad \bar{A}_1 - \bar{A}_3 = 12.78 - 7.11 = 5.67;$$

$$\bar{A}_1 - \bar{A}_4 = 12.78 - 8.78 = 4.00; \qquad \bar{A}_1 - \bar{A}_5 = 12.78 - 11.44 = 1.34.$$

Since the first two differences (involving a_2 and a_3) exceed the critical value of 4.35, we can conclude that the specific interfering activities represented by these two conditions produced a significant memory deficit. The other two experimental treatments (a_4 and a_5) did not result in a significant loss of memory.

The Tukey Test

The Tukey test is designed to maintain the *FW* rate at the chosen value of α_{FW} for the entire set of pairwise comparisons.[8] The test is performed easily by arranging the treatment means in ascending order of magnitude on the dependent variable, as illustrated in Table 8-3. Entries within the body of this table represent the differences between any two treatment means. In the first row, for example,

$$\bar{A}_2 - \bar{A}_3 = 7.89 - 7.11 = .78; \bar{A}_4 - \bar{A}_3 = 8.78 - 7.11 = 1.67; \text{and so on.}$$

Differences are not entered for comparisons below the main diagonal of the table, since the listing would be an exact mirror image of the differences appearing above the diagonal.

The next step is to calculate the *minimum* pairwise difference between means that must be exceeded to be significant with the Tukey test. This value ($\bar{d}_T$) is given by the formula

$$\bar{d}_T = \frac{q_T\sqrt{MS_{S/A}}}{\sqrt{s}}, \tag{8-8}$$

where q_T refers to an entry in the table of the studentized range statistic (Table [illegible] in Appendix A), $MS_{S/A}$ is the error term from the overall analysis of variance,

[8] Like the Scheffé test, was designed to cover all comparisons, but it is more powerful than the Scheffé for pairwise tests and less powerful than the Scheffé for complex contrasts.

Table 8-3 An Example of Pairwise Comparisons

	LEVELS (ORDERED BY SIZE OF TREATMENT MEANS)				
	a_3	a_2	a_4	a_5	a_1
MEANS	7.11	7.89	8.78	11.44	12.78
$\bar{A}_3 = 7.11$	–	.78	1.67	4.33	5.67
$\bar{A}_2 = 7.89$		–	.89	3.55	4.89
$\bar{A}_4 = 8.78$			–	2.66	4.00
$\bar{A}_5 = 11.44$				–	1.34
$\bar{A}_1 = 12.78$					–

and s is the sample size for each group. If you look at Table A-6, you will see that three quantities enter into the determination of q_T: df_{error} (the df associated with the $MS_{S/A}$), r (the number of treatment means—a in this design), and α_{FW} (the FW error rate chosen for the Tukey test).[9] For this example, $df_{\text{error}} = df_{S/A} = 40$, $r = a = 5$, and $\alpha_{FW} = .05$; the value of q_T is 4.04. Substituting in Eq. (8-8), we find

$$\bar{d}_T = \frac{(4.04)\sqrt{13.22}}{\sqrt{9}} = 4.90.$$

An inspection of the differences in Table 8-3 reveals that only the largest difference ($\bar{A}_1 - \bar{A}_3 = 12.78 - 7.11 = 5.67$) exceeds the critical value for the Tukey test and would be declared significant.

The Tukey test can be used in conjunction with the F test, although calculating $F_{\text{comp.}}$ for each pair is not as convenient as calculating mean differences. In any case, the critical value of F (F_T) against which $F_{\text{comp.}}$ is compared is given by the following formula:

$$F_T = \frac{(q_T)^2}{2}. \qquad (8\text{-}[illegible])$$

Sequential tests. Two popular alternatives to the Tukey test are the Duncan and the Newman-Keuls tests. These tests are distinguished by the fact that significance testing follows a series of sequential tests, each with a different critical value to establish the significance between pairs of means. The computational details of these tests may be found in Kirk (1968, pp. 91-94) and Winer (1971, pp. [illegible] 201). All three tests have been compared in Monte Carlo experiments designed to reveal how well they control FW error and what their power characteristics are, i.e., how well they detect treatment differences when they are present. The study most well known to psychologists was conducted by Petrinovich and Hardyck (1969); other relevant studies have been reported by Carmer and Swanson ([illegible]

[9]Table A-6 is also used with the Newman-Keuls test, described briefly in the next section, in which case r takes on different values depending on the particular pairwise difference under consideration.

and by Einot and Gabriel (1975). The collective evidence seems to support the conclusion that the Tukey test is preferred over the other two tests.[10]

The Fisher Test

A different sort of correction procedure is the protected least significant difference test, which I will call the Fisher test. The test consists of two steps: the test of the omnibus F, followed by the unrestricted testing of all pairwise differences if the overall F is significant. If the omnibus F is not significant, no additional tests are conducted. A formula for the critical value of the Fisher test ($\bar{d}_F$) is

$$\bar{d}_F = \frac{t\sqrt{2(MS_{S/A})}}{\sqrt{s}}, \tag{8-9}$$

where t is found in Table A-3 of Appendix A under the chosen value of α and the degrees of freedom associated with $MS_{S/A}$. (The F table can be used by taking the square root of the critical value of F under $df_{\text{num.}} = 1$ and $df_{\text{denom.}} = df_{S/A}$.) This critical value is equivalent to the criterion for an uncorrected *planned comparison* expressed in terms of the difference between two means. Transformed to an F,

$$F_F = F(1, df_{S/A}). \tag{8-9a}$$

To illustrate the calculations, I will again use the data from Table 8-2. The first step is to evaluate the overall F. From the treatment totals in Table 8-2 (p. 157), we find

$$[A] = \frac{\Sigma A^2}{s} = 4{,}356.89;$$

$$[T] = \frac{T^2}{(a)(s)} = 4{,}147.20;$$

$$SS_A = [A] - [T] = 209.69;$$

$$MS_A = \frac{SS_A}{df_A} = \frac{209.69}{5-1} = 52.42.$$

The ratio,

$$F = \frac{MS_A}{MS_{S/A}} = \frac{52.42}{13.22} = 3.97,$$

[10] Einot and Gabriel rejected the standard Newman-Keuls test because of its inadequate control of the error and studied instead a modified Newman-Keuls test that is too complicated for use by researchers. They recommend the Tukey test for its simplicity when compared with the modified Newman-Keuls test, and its favorable power characteristics relative to the sequential procedures compared in their study.

is significant, exceeding the critical value of $F(4, 40) = 2.61$ at $\alpha = .05$. With this outcome, the Fisher test permits us to evaluate all pairwise differences against the following critical difference:

$$\bar{d}_F = \frac{t\sqrt{2(MS_{S/A})}}{\sqrt{s}}$$

$$= \frac{(2.02)\sqrt{2(13.22)}}{\sqrt{9}} = 3.46.$$

Using this value to assess the differences presented in Table 8-3, we find five of the comparisons to be significant. (In contrast, the Tukey test, with a critical value of $\bar{d}_T = 4.90$, declared only one difference, $\bar{A}_1 - \bar{A}_3 = 5.67$, to be significant.)

You should note that the Fisher test involves no special corrections once the overall F is found to be significant. Familywise error is controlled in effect by conditionalizing one's decision to conduct pairwise comparisons on the significance of the omnibus F test. That is, type I errors can be committed only when this F is *significant*; and when the overall null hypothesis is true, this will happen only a small proportion of the time, namely, the proportion specified by α. Thus, *FW* error is kept under control on the average by reducing greatly the proportion of times when a researcher conducts pairwise comparisons and the overall null hypothesis is true.

Carmer and Swanson (1973) compared 10 post hoc correction techniques using Monte Carlo procedures and concluded that the Fisher test offers the sort of balance between type I error and power that most researchers would find attractive and acceptable. Cohen and Cohen (1975), for example, recommend the Fisher test as a general strategy for analyzing the outcome of a complex study (see pp. 162-165 of that text). What is captured by the Fisher test is the attitude that a significant ombibus F usually means that population treatment effects are probably present and that additional tests are used to identify them; a certain number of type I errors will occur in this second stage, but they are far outweighed by a need to detect the true differences.

A serious problem with the Fisher test is its apparent incompatibility with a planned-comparison approach. The primary function of post hoc test procedures is to protect against reporting too many "accidents," i.e., type I errors. The Fisher test performs this function by stopping the researcher 95 percent of the time when the null hypothesis is completely true, i.e., when there are no differences in the population. But consider a case in which a single "deviant" mean is added to a set of equal means. The Fisher test no longer offers the same sort of protection for these equal means that it provided originally. This is because the omnibus F will now be significant *more* than 5 percent of the time—on account of the presence of this single deviant mean—and there will be an increased "opportunity" to make type I errors on comparisons involving the original set of treatments. Since experiments with planned comparisons are designed to detect at least one difference that will be significant, a researcher is generally assured a significant omnibus F and, as a

[illegible] provided by the Fisher test to conduct uncorrected post hoc [illegible] reason, then, it seems inappropriate to use the Fisher test when [illegible] comparisons are involved in the design or "dependable" treatment [illegible] control-experimental effects, have been incorporated into the [illegible]

[illegible] problems with the Fisher test should be mentioned. First, the test [illegible] pairwise comparisons, but not all interesting research questions fit [illegible] It is probably safe to assume, however, that the Fisher test [illegible] complex comparisons are included in the analysis since the test [illegible] errors to the chosen level of significance regardless of what one [illegible] the Fisher test seems to run into serious difficulties when un- [illegible] paired with unequal sample sizes in an experiment (Keselman, [illegible], 1979).

[illegible], we need to know more about the properties of the Fisher test [illegible] widespread use advocated by Cohen and Cohen (1975). If my [illegible] compatibility of the planned-comparison approach and the Fisher [illegible] test can be recommended only for situations in which no specific [illegible] predicted by the researcher.

[illegible]ICAL POINTS TO CONSIDER

[illegible] the serious dilemma faced by all researchers. On the one [illegible] to design experiments that are analytically rich and lead to a [illegible] comparisons and analyses, while on the other hand, you fully [illegible] type I error is present whenever two or more statistical tests [illegible] the analysis of a single experiment. At one extreme you could [illegible] that all possible findings are important and resist either restricting [illegible] comparisons undertaken or employing some procedure to control [illegible] error. At the other extreme, you could maintain that accidental [illegible] avoided and adopt strong corrective methods to prevent the [illegible] error resulting from the assessment of two or more compari- [illegible] you would evaluate all comparisons at the same *PC* rate [illegible] the second case, you would subject all comparisons to the Scheffé [illegible] .05). The extreme positions are clear, but few individuals [illegible] point of view. What most of us do is to strike some balance be- [illegible] in an attempt to detect the presence of the most impor- [illegible] maintaining a reasonable control of *FW* error.

[illegible], I will discuss a number of ways to resolve the unavoidable [illegible] type I and type II errors. I will begin by considering a way to re- [illegible] large number of planned comparisons to a smaller and more [illegible] the outcome of certain key comparisons to guide the rest [illegible] I will mention a different sort of strategy that allows a [illegible] light control over *FW* error and avoids loss of power as

well. Finally, I will offer some general recommendations for dealing with the evaluation of multiple comparisons.

Conditionalized Planned Comparisons

An important stage in the design of any experiment is to consider the nature and quality of the information obtainable from the study as planned. At this point, we can verify that the study will permit answers to our original research questions. This listing also provides a form of early warning that the number of planned comparisons approaches or exceeds the number beyond which a researcher will begin to become concerned about familywise type I error. If this number is not exceeded, we will assume that the researcher will test these comparisons directly once the data are collected and make no adjustment for the theoretical size of the *FW* error. But what if the number of planned comparisons exceeds this value? One might decide to do nothing and to live with the estimated *FW* error. Alternatively, one could reduce the length of the list to include only comparisons of primary concern, in which case the "secondary" comparisons might be subjected to some form of *FW* control while the more important comparisons are not.

There is another way of dealing with this problem, which is best described by a concrete example. Suppose an experiment is conducted that compares the performance of subjects following the administration of two drugs, A and B. Suppose further that two different control conditions were deemed necessary and that two batches of drug A were available to the experimenter. The experiment can be viewed as a single-factor design with $a = 5$ treatment conditions. The design and 20 single-*df* comparisons that ask potentially meaningful questions of this experiment are presented in Table 8-4.

Although this set of meaningful comparisons falls far short of including the 90 single-*df* comparisons possible when $a = 5$, the number of comparisons probably exceeds the threshold beyond which most experimenters become concerned with *FW* error. Suppose, however, we test comparison 1 first and conditionalize the remaining 19 comparisons on the outcome of this single test. This particular comparison assesses the equivalence of the two control conditions. If the comparison is *not* significant, we will probably consider combining the two groups in any further comparisons and omit from consideration comparisons that involve the two control conditions separately. On the other hand, if the comparison is *significant*, we will be forced to evaluate drug effects with each control separately and probably omit from consideration comparisons where the two control conditions are combined.

This conditionalizing of subsequent comparisons on the outcome of comparison 1 is summarized in the upper portion of Table 8-5. Comparisons listed on the left are the ones conducted if the observed difference between the two control means is not significant, while comparisons listed on the right are the ones conducted if the difference is significant. In the first case, the total number of comparisons tested is 10, including the initial comparison; in the second case, the total

Table 8-4 A Set of Meaningful Comparisons

MEANINGFUL COMPARISONS	CONTROL 1	CONTROL 2	DRUG A_1	DRUG A_2	DRUG B
1	+1	−1	0	0	0
2	+3	+3	−2	−2	−2
3	+1	+1	0	0	−2
4	+1	+1	−1	−1	0
5	+3	0	−1	−1	−1
6	+1	0	0	0	−1
7	+2	0	−1	−1	0
8	0	+3	−1	−1	−1
9	0	+1	0	0	−1
10	0	+2	−1	−1	0
11	+1	+1	−2	0	0
12	+1	+1	0	−2	0
13	+1	0	−1	0	0
14	+1	0	0	−1	0
15	0	+1	−1	0	0
16	0	+1	0	−1	0
17	0	0	+1	−1	0
18	0	0	+1	+1	−2
19	0	0	+1	0	−1
20	0	0	0	+1	−1

number is 15. In either case, however, the total number of comparisons tested has been reduced from the original number of 20.

Comparison 17 can be used in a similar fashion to narrow down the number of planned comparisons actually tested in the analysis of this experiment. In this case, the comparison involves an assessment of the equivalency of the two batches of drug A. If the two batches produce equivalent results, we would probably combine the two groups in all further analyses involving drug A. The comparisons we would under these circumstances are listed on the left in the middle section of Table 8-5. If the two batches are *not* equivalent, we are forced to conduct separate analyses for the two batches and omit comparisons where they are combined. These comparisons are listed on the right in the table. The total number of comparisons tested including the initial one is 12 when comparison 17 is not significant and 13 when comparison 17 is significant. Again, there is a reduction in the number of comparisons tested when we use this conditionalized, or branching, technique.

If we conditionalize on *both* of these initial comparisons, we can reduce the number tested still further, as is illustrated in the fourfold table presented in the lower section of Table 8-5. The largest number of comparisons tested in this arrangement if we include the first two is 10, when both comparisons are significant, and the smallest number is 6, when neither comparison is significant.

This plan for reducing the number of planned comparisons would probably not create any controversy. Most researchers would agree that one should combine the two control conditions or the two batch conditions, since there is no com-

Table 8-5 An Example of Conditionalizing Planned Comparisons

TEST OF COMPARISON 1	
Not Significant	*Significant*
Test: comp. 2, 3, 4, 11, 12, 17, 18, 19, and 20.	Test: comp. 5, 6, 7, 8, 9, 10, 13, 14, 15, 16, 17, 18, 19, and 20.

TEST OF COMPARISON 17	
Not Significant	*Significant*
Test: comp. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, and 18.	Test: comp. 1, 3, 6, 9, 11, 12, 13, 14, 15, 16, 19, and 20.

	TEST OF COMPARISON 1	
TEST OF COMPARISON 17	*Not Significant*	*Significant*
Not Significant	Test: comp. 2, 3, 4, and 18.	Test: comp. 5, 6, 7, 8, 9, 10, and 18.
Significant	Test: comp. 3, 11, 12, 19, and 20.	Test: comp. 6, 9, 13, 14, 15, 16, 19, and 20.

pelling *theoretical reason* for not doing so. That is, the question is in essence *empirical,* rather than theoretical. Thus, the decision to combine or not to combine conditions can be reasonably decided by the outcome of the two statistical tests. The only question that might be raised concerns the presumed acceptance of the null hypothesis when the nonsignificant conditions are combined. One way to deal with this difficulty is to test the significance of these branching comparisons at a *higher* than normal probability level (e.g., $\alpha = .10$ or .25) in order to increase power and one's confidence in "accepting" the null hypothesis by combining nonsignificant conditions.

A Decision to Suspend Judgment

The major obstacle in the way of recommending a general plan for dealing with multiple comparisons that will satisfy most researchers is the differences in attitudes toward type I and type II errors held by different investigators. This stumbling block can be *circumvented* rather than removed simply by adding a *third*

decision category to the evaluation process when the two concerns responsible for the problem—*FW* and power—are in conflict. This conflict occurs, of course, whenever $F_{comp.}$ is significant as a planned comparison evaluated at an uncorrected α level of significance, but it is not significant when the α level is corrected for familywise type I error. Instead of deciding to reject or not to reject the null hypothesis in such ambiguous cases, I propose that we recognize this ambiguity by deciding to suspend judgment concerning the status of the null hypothesis. By taking no formal action in this situation, we avoid committing either a type I or a type II error. As a consequence, the decision contributes nothing to familywise error, since the null hypothesis has not been formally rejected. Suspending judgment calls attention to a potential true difference and avoids creating the obscurity often associated with a difference that is labeled "nonsignificant." The idea of suspending judgment is not new, but was suggested by Hays (1973, pp. 350-353) to deal with a similar problem in the evaluation of the overall *F* test. Applied to multiple comparisons, the use of a third decision category of suspending judgment introduces flexibility and clarity into a situation in which arbitrariness and ambiguity instead have been the rule.

How might this new procedure work out in practice? The first step is to determine the critical value of the test statistic at the *uncorrected* level of significance (α). This value, which I will call CV_α, is used to define the lower boundary of the rejection region for planned comparisons. The second step is to set a more stringent criterion for rejection that reflects one's concern for familywise type I error. This value, which I will call CV_{FW}, refers to the rejection region for the correction technique considered most appropriate for the analysis. I can now state the decision rules as follows:

If the test statistic equals or exceeds CV_{FW}, reject H_0.

If the test statistic falls between CV_α and CV_{FW}, suspend judgment.

If the test statistic is less than CV_α, do not reject H_0.

As an illustration, let's return to the experiment used as a numerical example of the Scheffé test (pp. 151-152). For that particular experiment, the critical value for the Scheffé test was $F_S = 10.44$, while the critical value for planned comparisons was $F_\alpha = 4.08$. The test statistic calculated from the data was an *F* ratio of [illegible]. The decision rules for this case become:

If $F_{comp.} \geq 10.44$, reject H_0.

If $F_{comp.}$ falls between 4.08 and 10.44, suspend judgment.

If $F_{comp.} < 4.08$, do not reject H_0.

[illegible] for $F_{comp.}$ in this example was 14.30 and H_0 is rejected.

As another illustration, consider the example presented in conjunction with the [illegible] test (pp. 155-156). In this case, the test statistic is expressed as the

difference between two means, symbolized as $\hat{\psi}$. For the Tukey test, the critical value is $\bar{d}_T$ = 4.90. The corresponding value for planned comparisons, calculated from Eq. (8-9) on p. 158, is 3.46. The decision rules are:

If $\hat{\psi} \geqslant 4.90$, reject H_0.

If $\hat{\psi}$ falls between 3.46 and 4.90, suspend judgment.

If $\hat{\psi} < 3.46$, do not reject H_0.

If you refer to Table 8-3 (p. 156), where all possible pairwise differences are enumerated, you can see that one of the comparisons exceeds 4.90 and leads to the rejection of H_0 (5.67), four fall between 3.46 and 4.90 and lead to the suspension of judgment (4.89, 4.33, 4.00, and 3.55), and the remaining five are less than 3.46 and lead to the nonrejection of H_0 (2.66, 1.67, 1.34, .89, and .78).

The three-decision system recognizes the ambiguity that exists when a comparison is significant under one criterion (as a planned comparison) but not significant under a more severe criterion (as a corrected post hoc comparison). By suspending judgment, we avoid committing either type of error, and simply conclude that the evidence is not sufficiently strong to justify either one of the usual conclusions. Since we have not rejected the null hypothesis, no type I error is committed, and the *FW* error is left unaffected by this decision. Since we have suspended judgment, no type II error is committed, and interesting and unexpected findings, often overlooked when a Scheffé test or other correction technique is used, can be assimilated into the interpretation of the experiment and perhaps earmarked for future replication and study.

Recommendations and Guidelines

You are now in a position to make up your own mind concerning the evaluation of analytical comparisons. The fact that there is little agreement among commentators writing in statistical books and articles concerning specific courses of action to be followed with multiple comparisons simply means that the issues are complex, and that no single solution can be offered to meet adequately the various needs of researchers. Consequently, you should view the situation not with despair and frustration, but rather with a realization that you can and *must* work the problem out for yourself. The first step is to understand the basic issue—which is simple really—that analytical comparisons, which are conducted in virtually every experiment we will consider in psychology, increase type I error in a predictable and inevitable fashion. Does this bother you? Your *degree* of concern, which is based on your attitude toward the relative importance of type I and type II errors for you and for your research field, will contribute greatly to your decision. Whatever plan you may adopt, however, you should make some attempt to estimate the degree of power under which you will be operating for comparisons of primary interest to

you. I will now offer some recommendations and guidelines that may assist you in deciding how to evaluate the analytical comparisons derived from your research.

Planned Comparisons. Planned comparisons are usually the motivating force behind an experiment. These comparisons are targeted from the start of the investigation and represent an interest in particular combinations of conditions—not in the overall experiment. Planned comparisons are examples of what Tukey (1977) calls **confirmatory statistical analysis**, where specific questions that can be confirmed or disconfirmed are tested in an experiment. In keeping with the special status accorded planned comparisons, I recommend using the uncorrected *PC* rate, α, to evaluate the relevant test statistics. If the number of planned comparisons exceeds the number of degrees of freedom associated with the overall treatment mean square, I suggest the use of the modified Bonferroni test (pp. 147-149) to maintain the *FW* error for planned comparisons at the level dictated by df_A, namely, $(a - 1)(\alpha)$. Planned comparisons need not be orthogonal, although orthogonal comparisons do provide an unambiguous allocation of the treatment variation.

Post Hoc Comparisons. In post hoc data analysis, the type of question asked shifts from "Is *this* difference significant?" which characterizes planned comparisons, to "*Which* differences are significant?" which characterizes post hoc comparisons. The concern is with the whole set of treatments, rather than particular combinations of conditions. The probability of finding significant differences by chance depends on the number of treatment conditions; hence, it makes sense to worry about the *FW* rate under these circumstances.

It is my opinion that post hoc comparisons should be subjected to a more stringent standard to guard against committing an unacceptably large number of type I errors. Just which correction technique you choose depends on the nature of the comparisons examined in the post hoc analysis. If the comparisons consist only of differences between a control condition and several experimental conditions, the Dunnett test (pp. 153-155) is appropriate. If various simple pairwise differences are considered, the Tukey test (pp. 155-156) is recommended. Finally, if more complex comparisons are involved, the Scheffé test (pp. 151-153) is the best choice.

To counteract any loss of power when post hoc corrections are applied, I recommend the use of the decision category to suspend judgment (pp. 162-164). By suspending judgment on post hoc differences that would be significant as planned comparisons but are not sufficiently large to be significant with a post hoc test, we can minimize the danger of missing small but interesting findings discovered during the course of systematically combing through the data. The question of power can then be dealt with later—by conducting future experiments that are specifically designed to study these post hoc findings for which judgment was suspended. In short, what Tukey (1977) calls **exploratory data analysis**, the uncovering of interesting and unexpected findings, often generates the planned comparisons studied and examined in subsequent experiments.

8.5 EXERCISES[11]

1. Table 8-4 (p. 161) lists 20 potentially meaningful single-*df* comparisons for a particular experiment. Translate the sets of coefficients defining these comparisons into a verbal statement describing the question asked by each comparison.
2. The bottom portion of Table 8-5 (p. 162) summarizes an analysis plan that depends on the joint outcome of two preliminary statistical tests.
   (a) Using the modified Bonferroni test, determine the values of α_{planned} to be used in each of the four alternative plans. Use $\alpha = .05$ as your starting point. (Be sure to include the two preliminary comparisons in your calculations.)
   (b) What is the critical value of F in each of these situations? Assume $s = 10$.
3. Assume that we have a control group and seven experimental groups, with $s = 16$ subjects for each group. The $MS_{S/A} = 28.75$. The totals for each group are given below:

C	E_1	E_2	E_3	E_4	E_5	E_6	E_7
289	270	241	279	191	213	205	198

   (a) Is the overall F significant?
   (b) Use Dunnett's test to determine which of the treatment means is significantly different from the mean of the control group. Use a two-tailed test at $\alpha = .05$.
   (c) Make the same set of comparisons with the Scheffé procedure, $\alpha = .05$. Do your conclusions change?
4. Suppose we have an experiment with independent groups of $s = 7$ subjects randomly assigned to each of 8 treatment conditions. The error term $MS_{S/A} = 58.65$. The treatment sums are given below:

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
316	333	307	373	398	227	123	436

   (a) Conduct the Fisher test on all pairwise differences, using $\alpha = .05$.
   (b) Evaluate the same pairwise differences with the Tukey test, $\alpha = .05$.
   (c) How would you modify your conclusions in part (b) if you adopted the three-decision procedure described in Section 8.4 (pp. 162-164)?

[11]The answers to these problems are found in Appendix B, beginning on p. 568.