groups differ from one another. In this case, the omnibus test should be viewed as a precursor to Scheffé's method. As discussed earlier, if the omnibus test is statistically significant, there is at least one contrast that will be statistically significant with Scheffé's method, namely a contrast whose coefficients are given by

$$c_j = n_j(\bar{Y}_j - \bar{Y}) \qquad \text{(18, repeated)}$$

Thus, a statistically significant omnibus test is a signal that it is worthwhile to search for significant contrasts. On the other hand, if the omnibus test is non-significant, searching for any significant contrast using Scheffé's method is pointless because none exists. Thus, the omnibus test serves a very definite purpose, but it serves this particular purpose only in conjunction with Scheffé's method.

As further assistance, Tables 5.16 and 5.17 summarize the procedural details for the Bonferroni, Tukey, and Scheffé procedures. Table 5.16 provides formulas for hypothesis testing, and Table 5.17 provides formulas for forming simultaneous confidence intervals. Both tables provide procedures to use when homogeneity of variance is assumed, as well as when it is not. Although

TABLE 5.16
TEST STATISTICS AND CRITICAL VALUES FOR MULTIPLE-COMPARISON PROCEDURES

	Test Statistic	*Critical Value*
Assuming Homogeneity of Variance		
Bonferroni	$(\hat{\psi})^2 \Big/ \left[MS_W \sum_{j=1}^{a}\left(c_j^2 / n_j\right)\right]$	$F_{.05/C;1,\,N-a}$
Tukey	$\dfrac{n_g n_h (\bar{Y}_g - \bar{Y}_h)^2}{(n_g + n_h) MS_w}$	$(q_{.05;a,\,N-a})^2/2$
Scheffé	$(\hat{\psi})^2 \Big/ \left[MS_W \sum_{j=1}^{a}\left(c_j^2 / n_j\right)\right]$	$(a-1)F_{.05;a-1,\,N-a}$
*Without Assuming Homogeneity of Variance**		
Bonferroni	$(\hat{\psi})^2 \Big/ \left[\sum_{j=1}^{a}\left(c_j^2 / n_j\right)s_j^2\right]$	$F_{.05/C;1,df}$
Tukey	$\dfrac{(\bar{Y}_g - \bar{Y}_h)^2}{\dfrac{s_g^2}{n_g} + \dfrac{s_h^2}{n_h}}$	large n: $(q_{.05,a\,df})^2/2$ small n: $V^2_{.05;a,df}$
Scheffé	$(\hat{\psi})^2 \Big/ \left[\sum_{j=1}^{a}\left(c_j^2 / n_j\right)s_j^2\right]$	$(a-1)F_{.05;a-1,\,df}$

*For all procedures, $df = \dfrac{\left(\sum_{j=1}^{a} c_j^2 s_j^2 / n_j\right)^2}{\sum_{j=1}^{a}\left(c_j^2 s_j^2 / n_j\right)^2 / (n_j - 1)}$

TABLE 5.17
FORMULAS FOR FORMING SIMULTANEOUS CONFIDENCE INTERVALS

Assuming Homogeneity of Variance	
Bonferroni	$\hat{\psi} \pm \sqrt{F_{.05/C;1,N-a}} \sqrt{MS_W \sum_{j=1}^{a}\left(c_j^2 / n_j\right)}$
Tukey	$(\bar{Y}_g - \bar{Y}_h) \pm (q_{.05;a,N-a} / \sqrt{2}) \sqrt{MS_W \left(\dfrac{1}{n_g} + \dfrac{1}{n_h}\right)}$
Scheffé	$\hat{\psi} \pm \sqrt{(a-1)F_{.05;a-1,N-a}} \sqrt{MS_W \sum_{j=1}^{a}\left(c_j^2 / n_j\right)}$
*Without Assuming Homogeneity of Variance**	
Bonferroni	$\hat{\psi} \pm \sqrt{F_{.05/C;1,df}} \sqrt{\sum_{j=1}^{a}\left[\left(c_j^2 / n_j\right)s_j^2\right]}$
Tukey	large n: $(\bar{Y}_g - \bar{Y}_h) \pm (q_{.05;a,df} / \sqrt{2}) \sqrt{\left(s_g^2 / n_g\right) + \left(s_h^2 / n_h\right)}$ small n: $(\bar{Y}_g - \bar{Y}_h) \pm V_{.05;a,df} \sqrt{\left(s_g^2 / n_g\right) + \left(s_h^2 / n_h\right)}$
Scheffé	$\hat{\psi} \pm \sqrt{(a-1)F_{.05;a-1,df}} \sqrt{\sum_{j=1}^{a}\left[\left(c_j^2 / n_j\right)s_j^2\right]}$

* For all procedure, $df = \dfrac{\left(\sum_{j=1}^{a} c_j^2 s_j^2 / n_j\right)^2}{\sum_{j=1}^{a}\left(c_j^2 s_j^2 / n_j\right)^2 / (n_j - 1)}$

the entries in the tables assume that α_{EW} has been set at .05, other values of α_{EW} could be substituted for .05.

In closing, we should mention that research on multiple-comparisons procedures is active in the field of statistics. Readers who are interested in more details are advised to consult Bretz, Hothorn, and Westfall (2011); Hochberg and Tamhane (1987); Hsu (1996); Toothaker (1991); Westfall, Tobias, and Wolfinger (2011); or Wilcox (1987a, 2012a, 2012b).

SUMMARY OF MAIN POINTS

Chapter 5 introduces the difference between per-comparison Type I error rate, denoted as α_{PC}, and experimentwise Type I error rate, denoted as α_{EW}. This chapter focuses on special methods that are needed when the goal is to control α_{EW} instead of to control α_{PC}. Once a decision has been made to control α_{EW}, further consideration is required to choose an appropriate method of achieving this control for the specific circumstance. One consideration is whether all comparisons of interest have been planned in advance of collecting the data. If so, the Bonferroni adjustment is usually most appropriate, unless the number of planned comparisons is quite large or if all pairs of groups are to be compared to each other. Statisticians have devoted a great deal of attention to

methods of controlling α_{EW} for conducting all pairwise comparisons, because researchers often want to know which groups differ from other groups. We generally recommend Tukey's method for conducting all pairwise comparisons. Neither Bonferroni nor Tukey is appropriate when interest includes complex comparisons chosen after having collected the data, in which case Scheffé's method is generally most appropriate.

IMPORTANT FORMULAS

Probability of at least one Type I error: Pr(at least one Type I error)
$$= 1 - \Pr(\text{no Type I errors}) = 1 - (1-\alpha)^C \quad \textit{for orthogonal contrasts} \tag{1}$$

General form of CI for a contrast: $\hat{\psi} \pm CV\sqrt{MS_W \sum_{j=1}^{a}\left(c_j^2 / n_j\right)}$ (3)

Bonferroni inequality: $1 - (1-\alpha)^C \le C\alpha$ (5)

Expected number of Type I errors: ENEPE $= C\alpha_{PC}$ (7)

Experimentwise Type I error rate: $\alpha_{EW} = \dfrac{\text{number of experiments with errors}}{\text{number of experiments}}$ (8)

Expected number of Type I errors: ENEPE $= \dfrac{\text{number of errors}}{\text{number of experiments}}$ (9)

F test allowing for unequal variances: $F = \dfrac{(\hat{\psi})^2}{\sum_{j=1}^{a}(c_j^2 / n_j)s_j^2}$ (10)

Bonferroni CI allowing unequal variances: $\hat{\psi} \pm \sqrt{F_{.05/C;1,df}}\sqrt{\sum_{j=1}^{a}\left[\left(c_j^2/n_j\right)s_j^2\right]}$ (11)

General form of CI for a contrast: estimate ± (critical value) (estimated standard error) (12)

Pairwise F allowing unequal variances: $F = \dfrac{(\bar{Y}_g - \bar{Y}_h)^2}{\frac{s_g^2}{n_g} + \frac{s_h^2}{n_h}}$ (13)

Pairwise df allowing unequal variances: $df = \dfrac{(s_g^2/n_g + s_h^2/n_h)^2}{s_g^4/n_g^2(n_g-1) + s_h^4/n_h^2(n_h-1)}$ (14)

Maximum F value for any contrast: $F_{maximum} = SS_{max}/MS_W$ (15)

Upper bound for SS_ψ: $SS_\psi \le SS_B$ (16)

Coefficients for SS_{max} contrast: $c_j = n_j(\bar{Y}_j - \bar{Y})$ (18)

Maximum SS for any contrast: $SS_{max} = SS_B$ (19)

Maximum F value for any contrast: $F_{maximum} = SS_B/MS_W$ (20)

Maximum F value for any contrast: $F_{maximum} = (a-1)MS_B/MS_W$ (21)

Scheffé Critical Value: $(a-1)\,F_{.05;a-1,N-a}$ (22)

ONLINE MATERIALS AVAILABLE AT *DESIGNINGEXPERIMENTS.COM*

Data. Data sets in a variety of formats are available for download.
Computing: SAS and SPSS. We provide SPSS syntax, SAS syntax, and SPSS point-and-click directions for some of the analyses discussed in the chapter.
Computing: R. We provide detailed R syntax in a tutorial type fashion that details how to replicate essentially all analyses discussed in the chapter.

EXERCISES

1. An investigator decides to test the following four contrasts in a five-group study:

	1	2	3	4	5
ψ_1	1	−1	0	0	0
ψ_2	0	0	1	−1	0
ψ_3	1	1	−1	−1	0
ψ_4	1	1	1	1	−4

Find the α_{EW} level if each contrast is tested with an α_{PC} level of .05.

*2. A researcher has conducted a five-group study. She plans to test the following pairwise comparisons: μ_1 versus μ_2, μ_2 versus μ_3, and μ_4 versus μ_5.
   a. What multiple-comparison procedure should be used to maintain the α_{EW} level at .05?
   b. What will the critical F value be for each contrast, if there are 13 participants per group?
   c. Suppose that after looking at the data, the researcher decides to replace the comparison of μ_2 versus μ_3 with a comparison of μ_3 versus μ_4. What multiple-comparisons procedure should be used to maintain the α_{EW} level at .05?
   d. What will the critical F value be in Part c if there are 13 subjects per group?
   e. What implications does the difference in critical values you found in Parts b and d have for revising planned comparisons after having examined the data?

*3. The following summary data are obtained in a four-group study, with 25 participants per group:

$\bar{Y}_1 = 52$	$\bar{Y}_2 = 46$	$\bar{Y}_3 = 51$	$\bar{Y}_4 = 54$
$s_1^2 = 96$	$s_2^2 = 112$	$s_3^2 = 94$	$s_4^2 = 98$

After examining the data, the experimenter decides to compare the means of Groups 2 and 4. He finds that the mean difference is non-significant using Scheffé's method.

a. Is he correct that this mean difference cannot be declared significant using Scheffé's method? (You can assume homogeneity of variance.)
b. Is there a better method available for testing this contrast that will maintain α_{EW} at .05, although the contrast was chosen post hoc? If so, can the contrast be declared significant with this method?

4. The experimenter in Exercise 3 has decided to supplement his hypothesis test comparing Groups 2 and 4 with a confidence interval.
   a. Use an appropriate method to form a 95% simultaneous confidence interval for the difference between Groups 2 and 4, where this specific comparison has been chosen from the larger set of all pairwise comparisons. You may assume homogeneity of variance.
   b. The experimenter argues that the interval in Part a could be formed using Equation 5.3 and setting *CV* equal to 1.99, because he is forming only this single interval. Do you agree? Why or why not?

*5. This problem asks you to reconsider the data from Exercise 13 in Chapter 4. The data are given here once again:

	1	*2*	*3*	*4*
	3	7	9	11
	4	5	2	7
	5	6	5	11
	5	5	9	7
	3	7	5	4
Mean	4	6	6	8
Var (i.e., s^2)	1	1	9	9

We assume that all pairwise comparisons are to be tested and that α_{EW} is to be maintained at .05. Although all comparisons are of potential interest, this exercise only requires you to consider two specific comparisons: Group 1 versus Group 2, and Group 3 versus Group 4.

a. Test the difference in the means of Groups 3 and 4, first using MS_W as the error term and then using a separate error term. How do the results compare?
b. Test the difference in the means of Groups 1 and 2, first using MS_W as the error term and then using a separate error term. How do the results compare?
c. Which error term do you think is more appropriate here? Why?
d. Researchers are sometimes reluctant to use separate error terms because they believe that doing so will lessen their opportunity to find a statistically significant result. Are they correct that using a separate error term will lower their power? Explain your answer.

6. This problem uses the same data as Exercise 5. However, we assume here that the goal now is to form confidence intervals instead of testing hypotheses. Assume that a confidence interval is to be formed for each pairwise comparison, but as in Exercise 5, this exercise only requires you to consider two specific comparisons: Group 1 versus Group 2, and Group 3 versus Group 4.

   a. Form a 95% simultaneous confidence interval for $\mu_3 - \mu_4$, first using MS_W as the error term and then using a separate error term. How do the results compare?
   b. Form a 95% simultaneous confidence interval for $\mu_1 - \mu_2$, first using MS_W as the error term and then using a separate error term. How do the results compare?
   c. Based on the respective confidence intervals, which error term do you think is more appropriate here? Why?
   d. Researchers are sometimes reluctant to use separate error terms because they believe that doing so will result in wider confidence intervals. Are they correct that using a separate error term will produce wider intervals? Explain your answer.

*7. A graduate student has conducted a four-group study in which he tested the following three planned comparisons:

	1	*2*	*3*	*4*
ψ_1	1	−1	0	0
ψ_2	.5	.5	−1	0
ψ_3	1/3	1/3	1/3	−1

The sums of squares for the three comparisons are 75, 175, and 125, respectively. The value of MS_W equals 25, and there were 11 participants in each group. The student's adviser wonders whether the omnibus F test of $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ would be statistically significant for these data. Can you help her?

   a. Is it possible to perform the test of the omnibus null hypothesis from the available information? If so, is the test significant? If it is not possible, explain why not.
   b. Find the observed F value for each of the planned comparisons tested by the student. Which, if any, are statistically significant with an α_{EW} level of .05?
   c. What relationship, if any, is there between the single observed F value of Part a and the three observed F values of Part b?

8. A researcher has conducted an experiment with six independent groups of 12 participants each. Although the omnibus F test was non-significant, he decided to use Scheffé's method of multiple comparisons. He claims that his calculations revealed that the average of the first three groups was significantly different from that of the last three. How would you respond to his claim?

9. A graduate student has designed a study in which she will have four independent groups of seven participants each. Parts a–h ask you to decide which multiple-comparison procedure (MCP) should be used to achieve maximal power while maintaining experimentwise alpha at .05. For each part, tell which MCP she should use and briefly justify your answer.
   a. The student plans to test all pairwise comparisons.
   b. The student decides after having looked at the data to test all pairwise comparisons.
   c. The student plans to test only four pairwise comparisons.
   d. The student decides after having looked at the data to test only four pairwise comparisons.
   e. The student plans to test seven planned comparisons.
   f. After having looked at the data, the student decides to test seven specific comparisons.
   g. The student plans to test 20 planned comparisons. (Hint: The critical t value for $\alpha_{PC} = .05/20$ is 3.376.)
   h. After having looked at the data, the student decides to test 20 specific comparisons.

10. The following data were obtained in a four-group study:

	1	*2*	*3*	*4*
	6	6	3	5
	5	9	7	3
	7	9	6	1
	5	4	3	4
	3	5	4	3
	4	6	7	5
Mean	5.0	6.5	5.0	3.5
Var (i.e., s^2)	2.0	4.3	3.6	2.3

a. Are the four group means significantly different from each other?
b. Suppose all pairwise comparisons were investigated. If the α_{EW} level is maintained at .05, is the difference between the means of Groups 2 and 4 significant? (You can assume homogeneity of variance).
c. How can you explain the results of Parts a and b? What general pattern of means is most likely to produce this type of result?
d. What does this example imply about the necessity of obtaining a statistically significant omnibus test before using Tukey's HSD method to test all pairwise comparisons?

*11. A professor has obtained the following data for a three-group between-subjects design:

Group	*Mean*	*SDs*
1	10	10.00
2	10	14.00
3	22	12.41

There were 11 participants per group (i.e., 33 participants in all).
a. The professor claims that he can reject the omnibus null hypothesis. Do you agree? Show your work.
b. Having allegedly found the three groups to be somewhat different, the professor uses Tukey's HSD method to test all pairwise comparisons. He claims that no differences were significant. Do you agree? Show your work.
c. On the basis of the results found in Parts a and b, the professor argues that the omnibus test is misleading. He concludes that he cannot state that there are any differences among these three groups. Do you agree? Why or why not?

12. This problem uses the same data as Exercise 11. Suppose that the first two groups are active treatment groups, whereas the third group is a placebo control group. Further suppose that the professor who collected these data wants to form two confidence intervals, one comparing the first treatment group to the control, and a second comparing the second treatment group to the control.
a. Because none of the comparisons of interest are complex, the professor uses Tukey's HSD as the basis for maintaining experimentwise alpha. What does the professor find when he forms intervals based on this approach?
b. A colleague suggests to the professor that he should use Bonferroni instead of Tukey to ensure simultaneous confidence here. Do you agree? Whether or not you agree, find the appropriate intervals based on the Bonferroni approach.

c. A student suggests to the professor that another option might be to use Dunnett's method to form his intervals. Find the appropriate intervals using Dunnett's method.
d. How do the intervals you found in Parts a–c compare to one another? Which method is best here? Why?

13. A graduate student used a four-group between-subject design for her thesis. She had $n = 11$ participants per group. Her sample means are $\overline{Y}_1 = 12$, $\overline{Y}_2 = 13$, $\overline{Y}_3 = 20$, and $\overline{Y}_4 = 19$. The value of MS_W was 55.
a. Should she reject an omnibus null hypothesis that $\mu_1 = \mu_2 = \mu_3 = \mu_4$? Show your work.
b. Based on her answer to Part a, she decides to investigate which groups are different. She decides to test all pairwise differences, assuming homogeneity of variance and using an appropriate method for controlling familywise error rate. Does she obtain any significant differences? Why or why not?
c. Her adviser asks her to compare the average of Groups 1 and 2 with the average of Groups 3 and 4, again controlling for familywise error rate. She argues in light of Part b that testing the complex comparison here is fruitless because tests of complex comparisons are more conservative then tests of pairwise comparisons. Is she correct? Show your work or explain your answer.
d. She has shown the results of Parts a–c to her adviser, who is thoroughly confused. He argues that according to the results she claims to have obtained, she has shown that $12(\overline{Y}_1)$ and $20(\overline{Y}_3)$ are not significantly different, but that 12.5 (the average of 12 and 13) and 19.5 (the average of 19 and 20) are significantly different, which is obviously absurd. Is his argument correct?
e. Approaching this apparent contradiction through confidence intervals may be illuminating. Form appropriate 95% simultaneous confidence intervals for the difference between Groups 1 and 3, as well as for the complex comparison of the average of Groups 1 and 2 versus Groups 3 and 4.
f. Which interval(s) in Part e contain zero? Which of the two intervals is centered farther from zero? Is this the interval that does not contain zero? Explain this pattern of results.

14. In an experiment with five independent groups (5 participants per group), the omnibus F value observed is 3.00, just barely significant at the .05 level. Noticing that the sample means are $\overline{Y}_1 = 10$, $\overline{Y}_2 = 10$, $\overline{Y}_3 = 15$, $\overline{Y}_4 = 20$, and $\overline{Y}_5 = 30$, it is decided to test the following post hoc comparison: $\psi = -7\mu_1 - 7\mu_2 - 2\mu_3 + 3\mu_4 + 13\mu_5$.
a. Find SS for this comparison. Show your work.
b. What will the observed F value for this comparison be? Why?
c. Will the result in Part b be significant using Scheffé's method? Why or why not?
d. What is the value of MS_W here?

15. Dr. S.Q. Skew performed an experiment involving four treatment groups with 16 participants per group. His research assistant performed an SPSS analysis of the data, but it did not answer all of Skew's questions. So far, Skew knows from this analysis that $SS_B = 864$ and $SS_W = 4{,}320$. He also knows that the observed F for the pairwise comparison of Groups 1 and 2 is equal to 1.000 and that the observed F for the pairwise comparison of Groups 3 and 4 is only 0.111 (i.e., literally 1/9). Because neither of these is significant, Skew wants to compare the average of the first two groups versus the average of the last two groups. Unfortunately, unbeknownst to Skew, his assistant has lost the data. Knowing that you are a statistical whiz, the assistant comes to you desperate for help. Your task is to test this third comparison for significance. Show your work. Also, assume that Skew chose this contrast after having examined the data.

16. The following data are from a completely randomized (between-subjects) design:

1	*2*	*3*
48	59	68
54	46	62
47	49	53
54	63	59
62	38	67
57	58	71

Five psychologists analyze this data set individually, each with different goals in mind. Your task is to duplicate the results obtained by each.

a. Psychologist #1 formulates three planned comparisons of interest: Group 1 versus 2, 1 versus 3, and 2 versus 3. Perform these planned comparisons, assuming homogeneity of variance.
b. Psychologist #2 has no a priori comparisons, so she first performs the omnibus test. Following this, all pairwise comparisons are tested for significance, assuming homogeneity of variance. Once again, provide observed and critical values.
c. Psychologist #3 differs from Psychologist #2 only in that he decides not to assume homogeneity of variance for testing the comparison (don't worry about this assumption for the omnibus test). Once again, provide observed and critical values.
d. Psychologist #4 differs from Psychologist #2 only in that she decides post hoc to test not only all pairwise comparisons but also the average of Groups 1 and 2 versus Group 3. Like Psychologist #2, she assumes homogeneity. Once again, provide observed and critical values.
e. Psychologist #5 performs the same tests as Psychologist #4. However, Psychologist #5 has planned to conduct these particular tests prior to examining the data. Homogeneity is assumed.
f. Finally, write a brief explanation (one to two paragraphs) of why the various psychologists did not all arrive at the same conclusions regarding group differences. You need not specify one approach as "best," but you should explain the patterns of findings for these data. Also, you need not discuss all findings in relationship to one another; instead, focus your attention on differences that emerge and the reasons for such differences.

17. This problem uses the same data as Exercise 16. Suppose that these data were collected with the specific goal of identifying the best treatment, where higher scores on the dependent variable are considered better.
    a. Assuming homogeneity of variance, use an appropriate method to form two-sided confidence intervals for the best treatment. Write a brief interpretation of your findings.
    b. Assuming homogeneity of variance, use an appropriate method to form one-sided confidence intervals for the best treatment. Write a brief interpretation of your findings.
    c. How do your results in Part b compare to your results in Part a? Is the difference you found for these data consistent with the general pattern of the difference between one-sided and two-sided intervals for the best treatment? Explain your answer.
18. A psychologist has tested eight independent hypotheses. She has decided she wants to control the false discovery rate (FDR) for this set of hypotheses. The eight p values she has obtained are as follows: .041, .022, .276, .010, .523, .003, .024, and .165.
    a. Which, if any, hypotheses can she reject using an FDR of .05? Show your work or explain your answer.
    b. Suppose she had decided that it was important to control the experimentwise alpha level at .05 for this set of hypotheses. Which, if, any hypotheses would she be able to reject from this perspective?
    c. Briefly explain types of situations where it might be justifiable to control FDR instead of α_{EW} at .05.

19. A psychologist has tested 10 independent hypotheses. He has decided to control the false discovery rate for this set of hypotheses at .05. The 10 p values he has obtained are as follows: .04, .15, .02, .31, .06, .63, .01, .03, .46, and .08. Which, if any, hypotheses can he reject controlling the FDR at .05? Show your work or explain your answer.
20. Chapter 3 presented data comparing five different therapies for heavy drinking. The dependent variable was a log-transformed version of number of drinks per week. The purpose of the current exercise is to explore pairwise comparisons among the groups. You may assume homogeneity of variance throughout this exercise.
    a. Use an appropriate method to control the experimentwise Type I error rate at .05 while comparing all pairs of means.
    b. Form a confidence interval for the difference in means between each pair of groups while maintaining 95% confidence for the entire collection of intervals.
21. This exercise uses the same heavy drinking data as Exercise 20. However, now suppose that the researcher who collected the data plans to compare not only all pairs of means but also the difference between the average of groups that receive Community Reinforcement Approach (CRA) without disulfiram to the average of Standard groups.
    a. What is the most appropriate method to answer the researcher's questions while controlling the experimentwise Type I error rate? Explain your answer.
    b. Use the method you specified in Part a to test the researcher's questions of interest.
    c. Use the method you specified in Part a to form confidence intervals with a simultaneous confidence level of 95%.
22. This exercise continues to use the heavy drinking data of Exercises 20 and 21. However, now suppose that the researcher who has collected these data wants to identify the best treatment based on the log-transformed measure of drinking.
    a. Which method is most appropriate if the researcher wants to be able to estimate the magnitude by which the best treatment is in fact best? Why?
    b. Use the method you specified in Part a to answer the question of which treatment(s) may plausibly be best based on these data.
    c. Which method is most appropriate if the researcher does not want to be able to estimate the magnitude by which the best treatment is in fact best? Why?
    d. Use the method you specified in Part c to answer the question of which treatment(s) may plausibly be best based on these data.
23. This is an exercise to test your ability to analyze data from a one-way between-subjects design and to report the results appropriately. Your general task is to write an abbreviated version of a "Results" section. However, your write-up need only concern itself with statistical details, and only minimal interpretation of the results is necessary. The data to be analyzed are those reported by Kroes et al. in a 2014 *Nature Neuroscience* article, which was summarized at the beginning of Chapter 4. You do not need to understand the details of this study, but a brief description was provided in Exercise 21 of Chapter 4. As in that exercise, we will pretend that several different members of the research team have access to the data, but take somewhat different approaches to the data analysis. (As you probably realize, we want you to see how various decisions can potentially affect the conclusions we reach about the data.) The current exercise is different from Exercise 21 of Chapter 4 because we now want you to consider what statistical method should be used to control the experimentwise Type I error rate for each psychologist.

    Unfortunately the team of psychologists has not reached agreement on how to analyze their data. Your task is to duplicate the results obtained by each psychologist, following the general instructions outlined earlier. In all cases, you should use the most appropriate technique available for answering each

psychologist's questions while also maintaining the experimentwise Type I error rate at .05 for that psychologist. Also, in all cases, you should provide justification for your conclusion as to whether a result is or is not statistically significant. In addition, supplement each hypothesis test you report with the corresponding confidence interval. In principle, it would be a good idea to investigate the extent to which scores are normally distributed, but for the purpose of this dataset you do not have to consider this assumption.

a. *Psychologist #1* argues that the theoretical hypothesis rests or falls on comparing the mean of Group A to the mean of Group B. Thus, this psychologist plans to test only this single comparison. He assumes homogeneity of variance across all 3 groups.
b. *Psychologist #2* is not entirely satisfied with the results reported to him by Psychologist #1, so he plans to test all pairwise comparisons among the groups. He assumes homogeneity of variance.
c. *Psychologist #3* is not entirely satisfied with the results obtained by either Psychologist #1 or #2. This psychologist asks you to suggest the best way to test all pairwise comparisons among the groups. What advice would you offer, and what happens if the data are analyzed this way?
d. Finally, write a brief explanation (three to four sentences) of why the three psychologists who analyzed these data differently did not all arrive at the same conclusions regarding group differences. You need not specify one approach as always "best," but you should explain the patterns of findings *for these data* in terms of general principles (e.g., are separate variance tests always less powerful than pooled variance tests?). More generally, you need not discuss all findings in relationship to one another; instead, focus your attention on differences that emerge, and the principles these differences illustrate.

24. The current exercise asks you to analyze the data from a study described in Chapter 3, Exercise 20 [James, E.L., Bonsall, M.B., Hoppitt, L., Tunbridge, E.M., Geddes, J.R., Milton, A.L., & Holmes, E.A. (2015). Computer game play reduces intrusive memories of experimental trauma via reconsolidation-update mechanisms. *Psychological Science*, *26*, 1201–1215]. As detailed in Chapter 3, the James et al. study is one of a series of studies by Emily Holmes and her colleagues that attempt to develop "a cognitive vaccine against traumatic flashbacks," by employing an innocuous computer game to lessen intrusive memory for traumatic events by disrupting the reconsolidation of memory for that event.

All 72 subjects viewed a 12-min trauma film consisting of 11 different incidents portraying actual or threatened death or serious injury, for example, a child being hit by a car or a man drowning. Twenty-four hours later, participants returned to the lab and were randomly assigned to one of four conditions: (1) a reactivation-plus-Tetris group, in which selected still images from all 11 trauma scenes were presented followed by playing the computer game Tetris for 12 minutes; (2) a no-task control group that was not given the memory-reactivation images nor asked to play Tetris but simply rated classical music excerpts for pleasantness and then sat quietly for the same length of time the first group was playing Tetris; (3) a Tetris-only group that did not see the selected still images but did play the computer game; and (4) a reactivation-only group that saw the selected still images but did not play Tetris. The investigators hypothesized that the memory of the film would be reactivated by the presented still images but that a taxing visuospatial task would create a capacity limitation that would interfere with reconsolidation of the traumatic memory, and hence lessen intrusive memories over the next week. Intrusive memories were defined for the participants as "scenes of the film that appeared spontaneously and unbidden in their mind" (James et al., 2015, p. 1204).

Over the next week, all participants completed daily diaries in which they were to mark when they experienced an intrusive memory (or to indicate they had not) and to write a description of the intrusive memory. The primary dependent variable of interest was the number of intrusive memories experienced over this 7-day period. (Several other dependent measures were collected which showed a similar pattern.) The raw data, as reported in the supplementary materials filed with the published

study, are available at *DesigningExperiments.com*. Assume a team of psychologists will be analyzing these data to test specific preplanned contrasts but disagree about how the tests should be carried out. All of these tests are motivated by the prediction that both reactivation of memory and a distracting task will be necessary for disruption of the reconsolidation of memory to occur, which disruption, the team agrees, is a theoretical mechanism that should result in a reduction of intrusive memories.

Although they don't agree on *how* the tests should be conducted, Psychologists 1, 2, and 3 at least agree on *which* comparisons should be tested, namely, five simple (i.e., pairwise) comparisons, three of which they expect to be significant and two to be non-significant. That is, they expect the Reactivation + Tetris condition to result in fewer intrusive memories than each of the other three conditions in turn, and they expect the No-Task control to not differ from either the Reactivation-only condition or the Tetris-only condition.

a. Psychologist #1 thinks that, given the robustness of F tests, it would be defensible to test the five contrasts assuming homogeneity of variance and normality, but controlling for experimentwise alpha at .05. What results would she obtain in her statistical tests?
b. Psychologist #2, in contrast to Psychologist #1, is concerned about heterogeneity of variance. That is, he wants to test the contrasts of interest without assuming homogeneity of variance across groups. However, after determining the omnibus test of a one-way ANOVA is significant, he argues that this means that he is justified in testing any contrast of interest at a per-comparison alpha of .05. What results would he obtain to his tests?
c. Psychologist #3 thinks it is appropriate both to allow for heterogeneity of variance and to control experimentwise alpha at .05. What tests should be done in this case and what results would be obtained?
d. Finally, write a brief explanation summarizing your conclusions and which tests of contrasts you believe should be reported. For any significant single-degree-of-freedom test results, report the mean difference, a confidence interval of the mean difference, a standardized mean difference, and a confidence interval around the standardized mean difference. Also explain the patterns of findings *for these data* in the context of general principles. For example, does allowing for separate variances always result in more or less powerful tests than assuming homogeneity of variance? Does controlling for experimentwise alpha always result in more or less powerful tests than not controlling for experimentwise alpha? Are Tukey-type procedures for pairwise comparisons always more powerful than Bonferroni-type procedures?

NOTES

1. The Expected Number of Errors per Experiment (ENEPE) is often referred to as the Error Rate per Experiment (ERPE). In fact, the first edition of our book used ERPE instead of ENEPE. However, in this edition we have chosen the term "Expected Number of Errors per Experiment" because we believe it more accurately describes the appropriate concept.
2. Bonferroni-adjusted confidence intervals for the pairwise contrasts can be obtained directly with both SAS and SPSS as long as we are willing to assume homogeneity. We have chosen to show hand calculations here for the sake of consistency, but in practice some work can be saved by relying on SAS or SPSS for the first three intervals shown in Table 5.5.
3. Tukey actually developed several multiple-comparisons procedures, which at times has resulted in confusing labels for the various techniques. The particular method we describe is referred to as Tukey's WSD (for Wholly Significant Difference), Tukey's HSD (for Honestly Significant Difference), or Tukey's T Procedure. As we will see later, the "wholly" and "honestly" terms serve to distinguish Tukey's method from Fisher's LSD (Least Significant Difference), which does not always properly control the α_{EW} level. Also, when we discuss within-subject designs (i.e., repeated measures designs) in Chapters 11–14, we

will see that the Bonferroni approach is better than Tukey's technique for testing pairwise comparisons of within-subject means.

4. Tukey originally developed a more general formula that allowed for tests of complex comparisons and pairwise comparisons, but Scheffé's procedure is more powerful for testing complex comparisons.
5. For our purposes, it suffices to state that the studentized maximum modulus distribution is similar in concept to the studentized range distribution. Readers seeking a more mathematical treatment are referred to Dunnett (1980) and to Hochberg and Tamhane (1987).
6. In most published tables of the studentized maximum modulus distribution, the columns refer to the number of comparisons being tested. We have chosen to present the columns in terms of the number of groups because we only discuss the distribution in the context of performing all pairwise comparisons.
7. Notice also that while we can assert that combination therapy is better than drug therapy, it is plausible that combination therapy is no better than diet and biofeedback while at the same time drug therapy is no worse than diet and biofeedback. In other words, we seem to have concluded that μ_1 and μ_4 differ from one another but neither is different from μ_2 and μ_3. A moment's reflection should convince you that if μ_1 and μ_4 differ from one another, they cannot both equal μ_2 (much less both μ_2 and μ_3). The explanation for this predicament is that there is more statistical power for detecting the difference between the largest and smallest population means than for intermediate means. There is no logical contradiction as long as we remember that a non-significant statistical test does not imply that the null hypothesis is exactly true. Another way of explaining this pattern of results is to rely on confidence intervals. For example, Table 5.7 shows us that the interval for $\mu_4 - \mu_2$ overlaps the interval for $\mu_1 - \mu_2$ (in particular, both intervals contain zero), but the mere fact that the intervals overlap does not imply that μ_4 equals μ_1.
8. Suppose that we define a contrast to have coefficients given by $c_j = n_j(\bar{Y}_j - \bar{Y})$. The sum of squares for this contrast will equal

$$SS_\psi = (\hat{\psi})^2 \Big/ \sum_{j=1}^{a} c_j^2/n_j$$

However, $\hat{\psi}$ is defined to be

$$\hat{\psi} = \sum_{j=1}^{a} c_j\bar{Y}_j = \sum_{j=1}^{a} n_j(\bar{Y}_{\cdot j} - \bar{Y})\bar{Y}_j$$

Substituting for $\hat{\psi}$ and c_j in the expression for SS_ψ yields

$$SS_\psi = \left[\sum_{j=1}^{a} n_j(\bar{Y}_j - \bar{Y})\bar{Y}_j\right]^2 \Big/ \left[\sum_{j=1}^{a} n_j^2(\bar{Y}_j - \bar{Y})^2/n_j\right]$$

which immediately reduces to

$$SS_\psi = \left[\sum_{j=1}^{a} n_j(\bar{Y}_j - \bar{Y})\bar{Y}_j\right]^2 \Big/ \sum_{j=1}^{a} n_j(\bar{Y}_j - \bar{Y})^2$$

It can be shown through some simple algebra that

$$\sum_{j=1}^{a} n_j(\bar{Y}_j - \bar{Y})\bar{Y}_j = \sum_{j=1}^{a} n_j(\bar{Y}_j - \bar{Y})^2$$

Making this substitution into the numerator of SS_ψ, we have

$$\begin{aligned} SS_\psi &= \left[\sum_{j=1}^{a} n_j(\bar{Y}_j - \bar{Y})^2\right]^2 \Big/ \sum_{j=1}^{a} n_j(\bar{Y}_j - \bar{Y})^2 \\ &= \sum_{j=1}^{a} n_j(\bar{Y}_j - \bar{Y})^2 \\ &= SS_B \end{aligned}$$

9. See Hochberg and Tamhane (1987) for a review of these studies. However, Kaiser and Bowden (1983) found that the Brown-Forsythe procedure can in some situations produce too many Type I errors. They propose multiplying the Brown-Forsythe critical value by the term $(1 + (a - 2)/df)$, where df is the denominator degrees of freedom.