Chapter 10

Contrasting Means in Between-Subjects Designs

10.1 OVERVIEW

In Chapter 8, we learned how to test the omnibus null hypothesis that the means of all populations sampled in an experiment are equal. Although a significant *F* indicates that not all of the population means are equal, it does not reveal the source of the differences among the means. For example, in the study of the effects of educational level on mean depression scores (see Table 8.5), we are left with questions such as: Does the mean for the population having only a high-school education differ from that for the population having some college education? Does the mean for the population with only a high-school education differ from the mean for a population with more than a high-school education; that is, from the mean of the other three populations? Within the context of a multi-factor designs we might ask these same questions, and also whether such differences depend on the sex of the subject.

Answering such questions requires calculating a test statistic and evaluating its significance. The calculations involve minor modifications of the *t* statistics presented in Chapter 6. Evaluating significance, however, is an issue when several tests are performed on a set of group means. Just as the probability of at least one head increases as the number of coin tosses increases, so does the probability of at least one Type 1 error increase as more significance tests are performed. The set of contrasts tested is referred to as a family, and the probability that the family contains at least one Type 1 error is referred to as the *familywise error rate*, or *FWE*. There are many factors that influence the *FWE* and many methods for controlling it.

The primary goals of this chapter are:

- To extend the *t* test for independent groups to testing contrasts of any form within a between-subjects design.
- To introduce a distinction between controlling the probability of a Type 1 error for an individual comparison, *EC* (for "error rate per comparison"), and controlling the probability of a Type 1 error within a family of comparisons, *FWE*.
- To distinguish among several different kinds of families of contrasts and present appropriate methods for controlling the *FWE* for each kind of family.

We first will develop these issues and applications in the context of a single-factor design. Once the procedures for testing contrasts are established, we will extend them to multi-factor designs, with emphasis on analyzing interactions.

10.2 DEFINITIONS AND EXAMPLES OF CONTRASTS

In Chapter 6, we reviewed procedures based on the *t* distribution for comparing a pair of means in the context of either testing hypotheses or constructing confidence intervals to estimate the difference between two means. In fact, the procedures introduced in Chapter 6 may be extended to more complex comparisons. In this chapter, we develop procedures for evaluating any type of contrast among means. We begin by defining a contrast.

A *contrast* of population means is denoted by the Greek letter *psi* (ψ) and is defined as a *linear combination* of the means; that is,

$$\psi = \sum_j w_j \mu_j \tag{10.1}$$

where μ_j denotes a population mean, w_j refers to a numerical weight, at least one w_j is not zero, and $\Sigma_j w_j = 0$.[1] To illustrate, consider the data set of Table 8.1. In this experiment, one group was taught to memorize words by the method of loci, a second group was told to form an image of each word, a third group was told to form a rhyme, and the fourth, a control group, was given no special instructions. We might wish to compare each of the strategies with the control condition. For example, the comparison of the control and image population means might be represented by

$$\psi_1 = \mu_{image} - \mu_{control} \tag{10.2a}$$

Rewriting Equation 10.2a to be explicit about the weights on the means and ordering the means according to their sequence in the data file gives

$$\psi_1 = (-1)\mu_{control} + (0)\mu_{loci} + (1)\mu_{image} + (0)\mu_{rhyme}$$

The control condition might be contrasted with the average of the three strategy conditions, in which case the contrast would be

$$\psi_2 = (1/3)(\mu_{loci} + \mu_{image} + \mu_{rhyme}) - \mu_{control} \tag{10.2b}$$

Again, being explicit about the weights on the means, Equation 10.2*b* may be written

$$\psi_2 = (-1)\mu_{control} + (1/3)\mu_{loci} + (1/3)\mu_{image} + (1/3)\mu_{rhyme}$$

Arguing that the image and loci strategies both involve imaging, the experimenter might ask whether their mean differs from the mean of the rhyme condition; then the contrast is

$$\psi_3 = (\tfrac{1}{2})(\mu_{loci} + \mu_{image}) - \mu_{rhyme} \tag{10.2c}$$

Again we may rewrite the contrast as

$$\psi_3 = (0)\mu_{control} + (\tfrac{1}{2})\mu_{loci} + (\tfrac{1}{2})\mu_{image} + (-1)\mu_{rhyme}$$

[1] The requirement that the weights sum to zero ensures that the contrasts deal with differences among means.

To construct a point estimate of a particular contrast on population means, $\hat{\psi}$, we simply substitute sample means for the corresponding population means:

$$\hat{\psi} = \sum_j w_j \overline{Y}_{.j} \tag{10.3}$$

For the contrasts in Equation 10.2, the point estimates are

$$\hat{\psi}_1 = (-1)(\overline{Y}_{control}) + (0)(\overline{Y}_{loci}) + (1)(\overline{Y}_{image}) + (0)(\overline{Y}_{rhyme}) \tag{10.4a}$$

$$\hat{\psi}_2 = (-1)(\overline{Y}_{control}) + (1/3)(\overline{Y}_{loci}) + (1/3)(\overline{Y}_{image}) + (1/3)(\overline{Y}_{rhyme}) \tag{10.4b}$$

$$\hat{\psi}_3 = (0)(\overline{Y}_{control}) + (1/2)(\overline{Y}_{loci}) + (1/2)(\overline{Y}_{image}) + (-1)(\overline{Y}_{rhyme}) \tag{10.4c}$$

Each of these equations estimates a different contrast of the four population means and therefore provides the basis for testing a different null hypothesis. Note that the zero weights in the first and third contrasts are not necessary for purposes of expressing the relevant contrast. However, it will be necessary to provide explicit weights for the means of all conditions when software is used to execute contrast computations, so it is good to get into the habit of providing weights for all conditions.

10.3 CALCULATIONS FOR HYPOTHESIS TESTS AND CONFIDENCE INTERVALS ON CONTRASTS

In Chapter 6, we presented calculations for the t statistic for two independent groups; that is, for *pairwise comparisons*. In this section we extend those calculations to contrasts involving more than two groups, such as those in Equations 10.4b and 10.4c. We also discuss the selection of weights when ns are unequal, and extend Welch's t' test of means when variances are heterogeneous to contrasts involving more than two groups.

10.3.1 Calculations When *ns* are Equal and Variances are Equal

A straightforward extension of the t test of Chapter 6 provides a test of contrasts in a one-factor design. We illustrate the test of contrasts using the data from the hypothetical memory experiment; those data were presented in Table 8.1 and are summarized in Table 10.1. The weights in the fourth row of the table are used for a test of the null hypothesis corresponding to Equation 10.2c:

$$H_0\text{: } (1/2)(\mu_{loci} + \mu_{image}) - (1)\mu_{rhyme} = 0$$

Table 10.1 Means and variances of the data in Table 8.1 with contrast weights

	Method			
	Control	Loci	Image	Rhyme
Mean	6.5	12.1	10.7	10.5
Variance	10.056	19.433	25.567	16.722
n	10	10	10	10
Weight	0	.5	.5	−1
Weight × 2	0	1	1	−2

Given the information in Table 10.1, we can now proceed to calculate the t statistic. The formula is

$$t = \hat{\psi} / s_{\hat{\psi}} = \frac{\sum_j w_j \overline{Y}_j}{\sqrt{MS_{S/A} \sum_j \frac{w_j^2}{n_j}}} \tag{10.5}$$

Note that the denominator involves $MS_{S/A}$, an average of all four within-group variances. Even though the control group mean is not included in the contrast, including the variance of the control group in the calculation of the error term is justified if we can assume homogeneous variances. The advantage of using $MS_{S/A}$ is that the estimate of error variance will be more accurate and our test will be more powerful because the error degrees of freedom are based on four groups rather than three.

We would not be justified in using $MS_{S/A}$ as the basis for computing our error term if the group variances differ, because doing so might bias the t test. In fact, the variances do not need to be very disparate before it makes sense to limit the error term to just the variances in the contrasted groups. For example, suppose $a = 3$, $n = 15$, and the group variances are 10, 12, and 29, respectively. Further, suppose the difference between the first two condition means is 2.5. If we test this difference using only the first two group variances, $F(1, 28) = 4.261$ and $p = .048$. However, if we assume homogeneous variances and use all three condition variances as the basis for our error term, the result is $F(1, 42) = 2.757$ and $p = 104$. Although this test is based on more *df*, the denominator of the F test is increased by the inclusion of the variance of the third group; consequently, the F ratio is smaller and nonsignificant. Even a nonsignificant result of a test of homogeneity of variance is not sufficient evidence to use $MS_{S/A}$ as our error term. In other words, we want to be quite confident of the assumption of homogeneity before we use $MS_{S/A}$ as the basis for our test. We will consider alternative analyses for the heterogeneous variance case in Section 10.3.4.

For the example of Table 10.1, we will assume homogeneity of variance. Therefore, given Equation 10.5 and the statistics in Table 10.1, we calculate

$$\hat{\psi} = (½)(12.1 + 10.7) - (1)(10.5) = 0.9$$

$$MS_{S/A} = (10.056 + 19.433 + 25.567 + 16.722)/4 = 17.944$$

and

$$\sqrt{\Sigma\, w_j^2/n_j} = \sqrt{1.5/10} = .3875 \quad \text{and} \quad \sqrt{MS_{S/A}} = \sqrt{17.944} = 4.236$$

Therefore,

$$\begin{aligned} t &= \hat{\psi} / s_{\hat{\psi}} \\ &= 0.9/[(4.236)(.3875)] \\ &= .548 \end{aligned}$$

The *df* associated with the t test are the *df* on which the estimate of the error variance is based; these are the *df* associated with $MS_{S/A}$, or 36. The choice of an appropriate critical value of t will be deferred until we discuss the concept of familywise error rate. However, it is clear that the computed t value of .548 is not significant by any reasonable criterion.

The computations illustrated for the hypothesis test might have been used to compute a confidence interval to estimate the contrast, $(1/2)(\mu_{loci} + \mu_{image}) - (1)\mu_{rhyme}$. In Chapter 6, we learned how to construct an interval estimate on the difference between two population means. The generalization of that procedure to any contrast is

$$\hat{\psi} \pm (t_{critical})(s_{\hat{\psi}}) \tag{10.6}$$

Again, discussion of the selection of a critical value of t will be deferred for now. The contrast and standard error are computed the same way as for the hypothesis test, and the standard error is, again, based on 36 *df* in our example.

We have been considering a contrast where some of the weights on the means are fractions (i.e., ½ and ½). When using software to do such computations, it is necessary to express fractional weights as decimals (e.g., .5). In some cases, converting fractions to decimals results in a repeating decimal place (e.g., 1/3 converts to .33333. . . .). To avoid the imprecision caused by arbitrarily truncating a repeating decimal, it is useful to multiply the weights in a contrast by a constant to convert the weights to integers. For example, multiplying the weights in our example by 2 converts the weights to 1, 1, and –2 (see row 4 of Table 10.1). Multiplying the weights of a contrast by a constant has no effect on the value of a t test because both the numerator and denominator of the t ratio are multiplied by the same constant. However, a change in the weights of a contrast will affect a confidence interval: The point estimate and the width of the interval will be multiplied by the constant. Therefore, the upper and lower bounds of a confidence interval should be returned to its original scale by dividing the two boundary values by the constant. The reader should compute a t test and confidence interval for our example contrast, using the weights 1, 1, and –2 to verify that the value of the t ratio is unaffected by the changed weights, whereas the confidence interval boundaries are multiplied by 2.

Most statistical software packages will perform the calculations illustrated in this section, and will do so even if group sizes or variances are not equal. For example, SPSS's *One-Way ANOVA* module (in the *Analyze* menu) has a *Contrasts* option that enables the user to select group weights. To take advantage of available software, the researcher must become comfortable with specifying the weights on each condition mean, and with expressing the weights as integers. Finally, if the interest is in constructing confidence intervals, it is important to remember to convert the bounds on each interval back to the original scale.

10.3.2 Weighting Means When *ns* Are Unequal: Equal-Sized Populations Assumed

Suppose we wish to compare the speeds of solving arithmetic problems in four different grades (e.g., Royer et al., 1999; see the *Royer RT_speed* data file on the book's website). Table 10.2 presents mean variances and class sizes (n). Although there are different numbers of students in the different grades, this is likely to be due to chance; that is, there is no reason to view the four sampled populations as unequal in size. If the various populations are equal in size, then any means that are averaged should receive equal weight in a contrast. For example, when testing the average speed of fifth-graders against the combined average of the sixth-, seventh-, and eighth-graders, the weights would be –1, 1/3, 1/3, 1/3 or, equivalently, –3, 1, 1, and 1. Equal weighting of means that are averaged on one side of a contrast usually will also be appropriate in the analysis of data from any true experiment in which the independent variable is manipulated. Using the integer weights, we can use Equation 10.5 to calculate the statistic for the contrast of fifth-graders' mean speed against that of the combined mean of the sixth-, seventh-, and eighth-graders. You should verify that $t = \hat{\psi}/s_{\hat{\psi}} = .680/.130 = 5.233$. Because the estimate of variability in the calculation of the standard

Table 10.2 Summary information for the Royer response time data

	Grade				
	5	6	7	8	
Mean	0.350	0.560	0.586	0.583	
Variance	0.033	0.028	0.031	0.038	
n	23	26	21	20	$N = 90$

error is $MS_{S/A}$, the *df* associated with the test are (90 – 4) or 86. Again, we defer selection of a critical value of t until our discussion of the *FWE*. Of course, we could also construct a confidence interval on the contrast. If we used the integer weights and resulting values of .680 for the contrast and .130 for the standard error, the upper and lower boundary values would be divided by 3 to return the interval to our original scale.

10.3.3 Weighting Means When *ns* Are Unequal: Unequal-Sized Populations Assumed

In many observational studies, differences in group sizes reflect differences in population sizes. A case in point is the *Seasons* study, which we cited previously. In Chapter 8, we tested the omnibus null hypothesis that mean depression scores were equal for four populations defined by their education level. The four groups were males with only a high-school education (*HS*), some college experience (*C*), a bachelor's degree (*B*), or graduate school experience (*GS*). Panel *a* of Table 10.3 presents group sizes, means, and variances.

Assume that one question of interest was whether the mean depression scores differed between males with a high-school education and all other males. Because the relative group sizes suggest that the *HS* population is considerably smaller than the others, we assume that the four populations vary in size. Then the mean of the last three populations, which we will denote as $\mu_{>HS}$ ("greater than high school"), would be a weighted average; that is,

$$\mu_{>HS} = \frac{w_C\mu_C + w_B\mu_B + w_{GS}\mu_{GS}}{w_C + w_B + w_{GS}}$$

and the null hypothesis of interest is

$$H_0\colon \mu_{HS} - \frac{w_C\mu_C + w_B\mu_B + w_{GS}\mu_{GS}}{w_C + w_B + w_{GS}} = 0$$

Because the t test of a contrast is not affected when all weights are multiplied by a constant, we can simplify things by multiplying the expression by $w_C + w_B + w_{GS}$, yielding the contrast

$$\psi = (w_C + w_B + w_{GS})\mu_{HS} - (w_C\mu_C + w_B\mu_B + w_{GS}\mu_{GS}) \tag{10.7}$$

and we can test the null hypothesis that this contrast equals zero.

Unless we know the actual sizes of the populations, we now need values of the *w*s. In many situations, the simplest and most reasonable will be the group sizes. Panel *c* of Table 10.3 presents output for Contrasts 1 and 2; the weights for the two contrasts are presented in Panel *b*. The weights

Table 10.3 Summary statistics (*a*), contrast coefficients (*b*), and test results (*c*) for depression scores as a function of educational level

(*a*) Statistics

Educational level	*n*	Mean	Variance
HS	19	6.903	34.541
C	33	3.674	5.97
B	37	3.331	9.861
GS	39	4.847	26.218

(*b*) Contrast coefficients

Contrast	Educational level			
	HS	*C*	*B*	*GS*
1	3	−1	−1	−1
2	109	−33	−37	−39

(*c*) Contrast test results (from SPSS)

	Contrast	Value of contrast	Std. error	*t*	*df*	Sig. (2-tailed)
Assumes equal variances	1	8.857	3.116	2.842	124	0.005
	2	318.922	113.204	2.817	124	0.006
Does not assume equal variances	1	8.857	4.181	2.118	20.527	0.047
	2	318.922	152.261	2.095	20.712	0.049

for Contrast 1 would be appropriate if we assumed that the populations are of equal size; that is not the case in this example, but Contrast 1 provides a comparison with Contrast 2, which uses weights based on the assumption that the group sizes reflect the population sizes. In Contrast 2, the weights on the means were calculated with Equation 10.7: the weight on the *HS* mean is the sum of the *n*s for the other three conditions; the weights on *C*, *B*, and *GS* are computed as −1 times the group size. Results are reported when equal variances are assumed and when they are not; we will discuss the calculations for the latter case in Section 10.3.4. For now, note that the assumption about variances can greatly influence the values of *t* and *p*.

Although the results for Contrasts 1 and 2 are very similar in this case, equal weighting and weighting by frequency can yield very different results; the distribution of group sizes is the critical factor. The reason results were similar in this example is that if we divide 109, −33, −37, and −39 by 36, we get 3.028, −.917, −1.028, and −1.083—not very different from 3, −1, −1, and −1.

We should point out that the issue of weights arises only with contrasts in which one or both subsets are based on at least two means. When testing pairwise comparisons (by far the most common situation), the weights will always be 1 and −1 for the two means involved in the comparison, and 0 for all other means.

10.3.4 Testing Contrasts When Variances Are Not Equal

In the situations we have considered to this point, $MS_{S/A}$ was used as our estimate of error variance in calculating the standard error for a contrast. However, there are two situations in which $MS_{S/A}$ is not an appropriate error term. In one case, the variances corresponding to the conditions involved in the contrast are very similar but different from the variances of those conditions not included in the contrast (that is, those having zero weight). The standard *t* is appropriate here, but the denominator should be based only on the variances corresponding to the included conditions. Degrees of freedom, due to the omitted group, are lost, but the standard error of the contrast is a valid denominator for the *t* test. For example, if there are three groups with variances 20, 21, and 5, and the means of the first two groups are to be compared, the variance of 5 should not be included in the denominator because the estimate of the standard error of the contrast will be too small, and the Type 1 error rate will be inflated.

The second situation is one in which there is heterogeneity of variance within the set of means that are to be contrasted. In this case, an extension of Welch's *t* test (*t*′; Welch, 1947; see Section 6.6.2) should be calculated. Recall that when variances are assumed to be homogeneous, the condition variances are pooled; however, when variances are assumed to be heterogeneous, the variances are not pooled and should be weighted differently according to the different weights on the corresponding means in the contrast. Also, the *df* are adjusted downward when unequal variances are assumed. The difference between the standard *t* results and those for *t*′ are illustrated in panel *c* of Table 10.3 where two sets of results are presented for each of the two contrasts of the *HS* depression mean with the mean of the other three groups. One result is obtained assuming equal variances, and the *t* is calculated as in Equation 10.5. The second result is obtained when equal variances are not assumed.

To obtain the result when equal variances are not assumed, calculate

$$\mathrm{t}' = \hat{\psi} \,/\, s_{\hat{\psi}} \tag{10.8}$$

where

$$s_{\hat{\psi}} = \sqrt{\sum_{j=1}^{a} \frac{w_j^2 s_j^2}{n_j}} \tag{10.9}$$

and the degrees of freedom are

$$df' = \frac{s_{\hat{\psi}}^4}{\sum_j \frac{w_j^4 s_j^4}{n_j^2(n_j - 1)}} \tag{10.10}$$

In summary, as can be seen in Table 10.3, the value of the test statistic depends both on whether the population variances are assumed to be equal and whether the populations are assumed to be equal in size.

10.4 EXTENDING COHEN'S *d* TO CONTRASTS

In Section 6.8, we introduced Cohen's *d*, a measure of the effect when there are two levels of the independent variable. This measure can be extended to contrasts in which several group means are involved. Assuming homogeneous variances, the general form of the standardized effect size for a contrast is

$$d = \hat{\psi}/\sqrt{MS_{S/A}} \tag{10.11}$$

Using the memory data summarized in Table 10.1, and the contrast illustrated there,

$$d = 1.8/\sqrt{17.944} = .42$$

Thus, although the *t* calculated for the contrast was quite small (.548), the standardized contrast is of medium size according to Cohen's (1988) guidelines. Without consideration of confidence bounds, it is difficult to evaluate this statistic, but it does leave open the possibility that power may have been lacking in the original test.

10.5 THE PROPER UNIT FOR THE CONTROL OF TYPE 1 ERROR

10.5.1 Defining a Family of Tests

As we stated in the Introduction, the probability of a Type 1 error increases with the number of significance tests. Therefore, if the probability of each significance test is set without regard to how many tests might be conducted, the error rate for the entire collection of tests may rise to an unacceptable level. Statisticians and researchers generally are agreed that the proper unit for control of the Type 1 error rate is not the individual test but a set of contrasts called a *family*. Before we address the question of how to limit the Type 1 error rate for the family, we should clarify the idea of a family of contrasts.

It is useful to distinguish between the *error rate per contrast* (*EC*)—the probability that a single contrast results in a Type 1 error—and the *familywise error rate* (*FWE*)—the probability that a set, or family, of contrasts will contain at least one Type 1 error. For a family of *K* independent tests,

$$\begin{aligned} FWE &= p(\text{at least one Type 1 error in the family}) \\ &= 1 - p(\text{no Type 1 errors in the family}) \\ &= 1 - p(\text{no Type 1 error on a single test})^K \end{aligned}$$

The probability of a Type 1 error on a single test is the *EC*. Therefore,

$$FWE = 1 - (1 - EC)^K \tag{10.12}$$

If a family consists of six independent tests each conducted at $EC = .05$, substitution in Equation 10.12 results in $FWE = 1 - (1 - .05)^6 = .265$; that is, even if the population means are all equal, the probability is .265 that one or more of the six tests will be significant. If the six tests are not independent, the exact value of *FWE* is difficult to calculate, but it is still greater than .05. In general, the larger the family, the more the *FWE* exceeds the *EC*. This suggests that in order to control *FWE*, we will need to adjust *EC* downward by an amount that will depend upon the size of the family of comparisons. This line of reasoning requires that we decide how to specify a family of contrasts. We will consider three alternatives.

An investigator working in a research area over a period of years might perform hundreds of experiments and test thousands of hypotheses. We might consider these thousands of tests to form a single family and set *FWE* equal to .05; however, this is not a reasonable specification of a family of tests, in part because it would result in an *EC* that would be infinitesimally small. Although this ultraconservative approach would result in a very low Type 1 error rate, the Type 2 error rate would soar to unacceptable levels. The experimenter could be confident that significant results revealed real effects but would miss finding many real effects. Because lowering the *EC* results in a reduction of power, the definition of family must be based on a compromise between concerns about Type 1 and Type 2 errors.

A more reasonable choice for a family of comparisons would be all of the tests conducted to analyze the results of an experiment. However, this approach to specifying families of comparisons is arbitrary because experiments may differ widely in the number of conditions they include, and thus the number of tests that might be performed. A researcher who performs simple experiments with few factors would perform fewer tests than a researcher who performs more complex, multifactor experiments with many conditions. This brings us to a third and more reasonable approach to specifying families of comparisons; namely, to identify families of tests with sources of variance in an experimental design. For example, in an experiment with three factors, a set of tests to understand the *AB* interaction would constitute one family of comparisons, and a set of tests to understand the main effect of *C* would constitute another. *FWE* would be controlled independently for the two families.

Identifying a family with the set of comparisons conducted to analyze a single source of variance seems a reasonable approach to defining families of contrasts in a few senses. First, it strikes a balance between control of Type 1 errors and power considerations by keeping the size of the family manageable. Related to this, compared to the two previously considered definitions of a family, this approach results in more consistency in the size of a family. Finally, there is a clear, substantive basis for the definition of a family when a family is identified with a source of variance.

Although the identification of a family of comparisons with a source of variance does a good job of addressing some important issues associated with controlling *FWE*, a researcher generally has many decisions to make regarding which contrasts to perform when analyzing a source of variance. For example, suppose we are comparing the effects of four different drugs on depression. There are six possible pairwise comparisons. In addition, we could compare each one of the drugs with the average of the other three, and with the average of two of the other three. That leads to a total of 22 possible significance tests. Many of the tests may be of interest, but we would suffer a substantial loss of power for each test if we conducted such a large set of tests while controlling *FWE* at a reasonable level. In short, the specification of a family of contrasts should involve a compromise between wanting to gain as much information as possible, and keeping the number of contrasts low enough to control *FWE* while still having reasonable power. In view of this, we should think hard about which hypotheses are of interest before we collect the data. We need to focus both the research design and the power of our significance tests on those questions that are of most interest to us.

10.5.2 Different Types of Families of Tests

In the sections that follow, we will describe several different methods for controlling *FWE*. The reason for the different methods has to do with the goals of procedures for controlling *FWE*. Any procedure for controlling familywise error rate attempts to (1) control *FWE* at a specified level (e.g., .05 or .10), while (2) maintaining good power to test individual contrasts. To meet these goals, a method must take into account both the size of the set of contrasts and the relations among those contrasts (e.g., independent or correlated). It is therefore useful to distinguish five categories of families of contrasts that differ in both these respects; these distinctions will help to organize the different procedures for controlling *FWE*.

1. *Planned contrasts.* A sensible research strategy is to include only those conditions in an experiment that are of interest, and to plan the contrasts to be tested before collecting the data. Focusing both the design and planned tests on only those contrasts that are important conserves time and effort in data collection and, by limiting the size of the family of contrasts, ensures more powerful tests of the contrasts when the familywise error rate is controlled.
2. *All pairwise comparisons.* Not all research can be planned to the point of specifying a set of contrasts *a priori*. A researcher must therefore be able to conduct tests based on an examination of the patterns of means that are actually observed. A simple, principled approach to analyzing a source of variance is to conduct the $(1/2)(a)(a - 1)$ tests of differences between all pairs of group means to try to summarize patterns of effects.
3. *All comparisons with a control condition.* A relatively common experimental design includes a single control condition that serves as a baseline for evaluation of several experimental conditions. Assuming a conditions including the control, there will be $a - 1$ significance tests of interest.
4. *Post hoc contrasts.* Despite thoughtful planning of contrasts, unanticipated differences may appear between means that may require more complex tests than simple pairwise comparisons. For example, it may appear that the mean across three experimental conditions is greater than the mean of a control condition. A researcher would, of course, want to explore those differences. However, the *FWE* is considerably larger in this situation than in the three preceding situations because the size of the family of tests is considerably larger, as we shall see.
5. *Contrasts when the independent variable is quantitative.* When the independent variable is defined by amount rather than type, the shape of the functional relation between the independent and dependent variable is usually of most interest. We will address the topic of trend analysis in Chapter 11.

To recap, the five categories of families differ in size and in the relations among the contrasts that comprise them. Thus, different methods for controlling *FWE* are appropriate for these different types of families. It will be useful to keep in mind that the same *t* test and confidence interval calculations may be applied in every case that we will consider. All that changes from one situation to another is the method of determining a critical value of *t*.

10.6 CONTROLLING THE *FWE* FOR FAMILIES OF *K* PLANNED CONTRASTS USING METHODS BASED ON THE BONFERRONI INEQUALITY

We begin with some very general methods that are applicable whenever several tests are planned. These methods apply not only to tests of differences among means, but also to tests of hypotheses about other parameters, such as proportions or correlations. *Note that it is not necessary that the omnibus F be significant prior to testing planned contrasts with the methods described in this section.*

In fact, power is lost by requiring a significant *F* before carrying out planned tests with these methods. What is critical is that the contrasts are decided on before the data are collected and a method for evaluating significance of the tests is used that maintains the familywise error rate at or below a reasonable level, presumably .05 or .10.

Equation 10.12 describes the relation between *FWE* and *EC* when the *K* tests are independent. Because this condition rarely holds, a more general statement of the relation is

$$FWE \le 1 - (1 - EC)^K \qquad (10.13)$$

In other words, the *FWE* is equal to *or less than* the term on the right with the inequality holding when the tests are not independent. Furthermore, if *K* tests are conducted with error rates EC_1, $EC_2, \ldots, EC_K$,

$$FWE \le \sum_K EC_K \qquad (10.14)$$

where EC_K is the probability of a Type 1 error for the K^{th} contrast. The relationship expressed in Equation 10.14 is known as the *Bonferroni inequality*, and it is the basis for several procedures for testing planned contrasts. From the inequality, it follows that if each of the *K* contrasts that make up the family is tested at $EC = FWE/K$, the probability of a Type 1 error for the family cannot exceed the *FWE*. If, for example, the family contains five planned contrasts, *FWE* will not be larger than .05 if each contrast in the family is tested at the .01 level.

In order to illustrate methods based on the Bonferroni inequality, we reconsider the memory experiment results summarized in Table 10.1. Table 10.4 contains an analysis of variance of those data, including results of tests of the three contrasts we considered earlier. The significance values for the three contrasts reported by SPSS are the *EC*s. According to this criterion, and assuming $\alpha = .05$, recall is significantly better in the image than in the control condition (Contrast 1), and the average of the three experimental methods yields significantly better recall than the control (Contrast 2). However, the reported *p*-values do not take into consideration the fact that three tests were performed on the means. Therefore, we will consider how each of two methods controls the *FWE* for this set of three tests.

10.6.1 The Dunn–Bonferroni Method (Dunn, 1961)

The Dunn–Bonferroni method for controlling *FWE* follows from Equation 10.14. If there are *K* contrasts, the *FWE* will not exceed a nominal value if the *EC* is set at that value divided by *K*. For example, assuming an *FWE* of .05, we test the three contrasts in Table 10.1 at the .0167 (.05/3) alpha level. The *t* statistics are those reported in Table 10.4; Equation 10.5 provides the formula for the *t* when variances are assumed equal, and Equations 10.8–10.10 provide the formulas for the *t* and *df* when variances are not assumed equal. If the exact *p*-value is available, we can control the *FWE* by comparing *p* with *FWE*/*K* and rejecting H_0 only for those contrasts where $p < FWE/K$. Looking at the SPSS output of Table 10.4 for our example, we would evaluate each of our three contrasts at the .0167 level. By this criterion, the contrast between the control and image conditions (i.e., Contrast 1) is no longer significant because the two-tailed *EC* is .033 (assuming homogeneity of variance).[2]

[2] Šidák (1967) proposed that, H_{0k}: $\psi_k = 0$ be rejected if $p_k \le 1 - (1 - FWE)^{1/K}$. Because $1 - (1 - FWE)^{1/K} > FWE/K$, this method has more power and a narrower confidence interval than the original Dunn–Bonferroni procedure. However, the difference is very small.

Table 10.4 Output for tests of three contrasts of means in a memory experiment (from SPSS)

(*a*) ANOVA

	Sum of squares	*df*	Mean square	*F*	Sig.
Between groups	173.900	3	57.967	3.230	.034
Within groups	646.000	36	17.944		
Total	819.900	39			

(*b*) Contrast coefficients

	Method			
Contrast	Control	Loci	Image	Rhyme
1	−1	0	1	0
2	−3	1	1	1
3	0	1	1	−2

(*c*) Contrast tests

	Contrast	Value of contrast	Std. error	*t*	*df*	Sig. (2-tailed)
Assumes equal variances	1	4.20	1.894	2.217	36	.033
	2	13.80	4.640	2.974	36	.005
	3	1.80	3.281	.549	36	.587
Does not assume equal variances	1	4.20	1.887	2.225	15.131	.042
	2	13.80	3.902	3.537	21.949	.002
	3	1.80	3.345	.538	20.466	.596

Confidence intervals based on the *FWE* have a form similar to those we encountered in Chapter 6; in general,

$$CI = \hat{\psi} \pm t_{FWE/K} s_{\hat{\psi}} \tag{10.15}$$

Consider Contrast 2 in Table 10.4. Inserting the means from Table 10.1, we have

$$\hat{\psi}_2 = (-3)(6.5) + (1)(12.1) + (1)(10.7) + (1)(10.5) = 13.80$$

Assuming equal variances, the standard error of this contrast is given in Table 10.4 as 4.640; it is calculated as in the denominator of Equation 10.5:

$$s_{\hat{\psi}} = \sqrt{MS_{S/A} \sum_j w_j^2/n}$$

The critical value of *t*, t_{FWE}, can be obtained from statistical packages or by interpolation from Appendix Table C.7. For a two-tailed *p* equal to .0167, the critical $t = 2.51$ when $df = 36$.

Substituting values into Equation 10.14, the confidence bounds for ψ_2 are

$$CI = 13.80 \pm (2.51)(4.64) = 2.154, 25.446$$

We are not quite done. Recall that our original coefficients were 1, −1/3, −1/3, and −1/3; we multiplied by 3 to have integer entries in SPSS. We return to our original scale by dividing the interval bounds by 3. The final bounds are .718 and 8.482. We would follow the same procedure for constructing confidence intervals for the other two contrasts in our example.

When confidence intervals are based on the *FWE*, they are interpreted somewhat differently than when they are based on the *EC*. It may help to understand the distinction if we assume many random replications of the memory experiment. In each replication, a set of three confidence intervals, one for each of the planned contrasts in Table 10.1, is calculated. We expect that in 95% of the replications, all three intervals will contain the true (population) value of the contrast. Thus, we are 95% confident that all of the confidence intervals in the family are accurate statements; that is, we are 95% confident that all three intervals contain the value of the population parameter being estimated. These intervals based on the *FWE* are referred to as *simultaneous confidence intervals*.

10.6.2 Hochberg's Sequential Method (1988)

Sequential methods (also referred to as stepwise, or multistage) test contrasts in several stages. Some of these methods such as Duncan's (1955) and Newman–Keuls (Keuls, 1952; Newman, 1932) fail to adequately control the Type 1 error rate and therefore will not be considered. A limitation of all sequential methods is that confidence limits cannot be calculated. However, sequential methods confer a power advantage over the Dunn–Bonferroni procedure if the researcher is solely interested in a set of hypothesis tests.

There are several sequential methods for controlling *FWE* for a family of planned hypothesis tests. The simplest of these are Holm's sequentially rejective method (1979), and Hochberg's step-up method (1988). The Hochberg method is the more powerful of the two; it is described in Box 10.1. The power of this procedure can be increased, but at the cost of greater complexity (Hommel, 1988; Rom, 1990).

Box 10.1 Hochberg's (1988) Sequential Testing Method

1. Rank order the *K* contrasts according to their *p*-values with p_1 being the smallest and p_K being the largest. In the example of Table 10.4, the contrasts would be ordered: ψ_2, ψ_1, and ψ_3.
2. If $p_K \le FWE$, all *K* null hypotheses are rejected. If not, consider the next largest *p*-value, p_{K-1}. If $p_{K-1} \le FWE/2$, reject the null hypothesis corresponding to the $K-1$ contrast and all remaining contrasts. If this test is not significant, test whether $p_{K-2} \le FWE/3$, and so on.
3. Using the example of Table 10.4 and assuming $FWE = .05$, we first compare .587 with .05. This fails and we compare the next largest *p*-value, .033, against .05/2 or .025. This fails and we compare the last *p*-value, .005, against .05/3, or .0167. This is the only significant contrast, so we can reject only H_0: $\psi_2 = 0$.

To summarize, the choice between the Dunn–Bonferroni and the Hochberg procedures depends on whether the researcher wants confidence intervals. The Hochberg method is more powerful when conducting hypothesis tests, but does not yield confidence intervals.[3]

[3] Both the Šidák and Hochberg methods are available in several statistical packages, including SPSS and Systat, but the results are based on comparisons of members of all pairs of group means.

10.7 TESTING ALL PAIRWISE CONTRASTS

10.7.1 The Studentized Range Statistic

We present several methods for controlling *FWE* across a family of pairwise comparisons; all of the methods are based on *q, the Studentized range statistic*. This statistic is the range of a set of observations from a normally distributed population, divided by the estimated standard deviation of the population. If the observations are group means,

$$q = \frac{\bar{Y}_{\max} - \bar{Y}_{\min}}{s_{\bar{Y}}} \qquad (10.16)$$

where $\bar{Y}_{\max}$ and $\bar{Y}_{\min}$ are the largest and smallest means in a set of a ordered means and $s_{\bar{Y}}$, the standard error of the mean, is

$$s_{\bar{Y}} = \sqrt{MS_e / n} \qquad (10.17)$$

assuming homogeneity of variance and equal *n*s. Critical values of q can be found in Appendix Table C.9 as a function of the *FWE*, a (the number of means), and the *df* associated with the error mean square. Harter, Clemm, and Guthrie (1959) provide a more extensive table.[4]

The Studentized range statistic is closely related to t; the two statistics differ only in their denominators, so it is a simple matter to derive the relationship between t and q. If we were to carry out a t test of the difference between the largest and smallest means, assuming equal variances and group sizes in a one-factor design, the statistic would be

$$t = \frac{\bar{Y}_{\max} - \bar{Y}_{\min}}{\sqrt{MS_{S/A}\,(2/n)}} \qquad (10.18)$$
$$= q/\sqrt{2}$$

Making explicit the relation between t and q makes it clear that the difference between the two classes of procedures is in the criterion for evaluating contrasts, rather than in the procedures for computing contrasts. Further, the relationship can be useful in comparing the results of procedures based on t (such as the Dunn–Bonferroni) with those based on q, and has also been the basis for dealing with unequal *n*s and unequal variances.

10.7.2 Tukey's (1953) *HSD* Test

Tukey's *HSD* (honestly significant difference) test controls the *FWE* for the set of all possible pairwise comparisons. It is a simultaneous method for testing hypotheses or constructing confidence intervals, meaning that a single critical value is used to evaluate all contrasts in a set. If the procedure is carried out without the help of software, it is helpful to compare differences against a critical difference. Specifically, a critical value of q is selected from Appendix Table C.9 and that value is multiplied by the standard error of the mean to find the critical difference between the two means that must be exceeded for significance. Once the critical difference is computed, it is a simple matter to evaluate any difference among group means. If the differences are ordered from largest to smallest, testing may stop once a nonsignificant difference is found because the remaining comparisons must also be nonsignificant.

A complete example of the procedure is presented in Box 10.2 using the summary statistics for the memory study that were presented in Table 10.1. Box 10.2 presents the necessary steps in testing all pairwise contrasts and for constructing simultaneous confidence intervals. The test is available in several statistical packages. For example, SPSS's *Post Hoc/Tukey* option reports the same bounds as in Box 10.2, and an exact *p*-value of .027 for the control versus loci difference.

[4] For significance levels other than those in Appendix Table C.9, users of SPSS can select the *Transform* option from the main menu, then *Compute Variable*; double-click on the *cdf.srange* option. Then, insert values for q (perhaps several trial values), a, and the error *df*. For example, in the left-hand panel, you may have a variable labeled p and in the right-hand panel, 2*(1 – CDF.SRANGE(5.05,4,19)). The p column of your data form should now show the value .02.

Box 10.2 Applying Tukey's *HSD* Test to the Memory Data of Table 10.1

Hypothesis tests

1. Order the means in Table 10.1 from smallest to largest:

Method:	Control	Rhyme	Image	Loci
Mean:	6.5	10.5	10.7	12.1

2. From Appendix Table C.9, find the value of q required for significance when $FWE = .05$, $a = 4$, $n = 10$, and $df = a(n - 1) = 36$. That value is approximately 3.81.
3. Calculate the standard error of the mean using the values of the variance in Table 10.1. Averaging the variance (assuming equal n), $MS_{S/A} = 17.944$, and

$$s_{\bar{Y}} = \sqrt{MS_{S/A}/n} = 1.340$$

4. We can now calculate a critical difference between means as $d_{crit} = SEM \times q_{crit} = 1.34 \times 3.81 = 5.1054$. All pairwise differences greater than this difference will be judged significant. For example, the largest difference (between the control and loci means) is 5.6 and is therefore significant. No other difference is larger than 5.1054, so this is the only significant difference when the *FWE* is controlled.

Confidence intervals

1. To construct confidence intervals, find the critical value of q and the standard error of the mean as above.
2. For any particular comparison of conditions, compute the difference between the means and compute an interval by: $\hat{\psi} \pm q_{(1-\text{confidence}),a,df}\, s_{\bar{Y}}$. For example, the 95% confidence limits on the difference in mean recall of the control and loci populations are:

$$5.6 \pm (3.81)(1.34) = .495,\ 10.705$$

In many studies, only a few of the possible pairwise comparisons will be of interest. In such cases, is the Tukey test or the Dunn–Bonferroni procedure the better choice for controlling *FWE*? Dunn (1961) has demonstrated that when all possible pairwise comparisons are tested, the Tukey procedure has the narrower confidence interval and is the more powerful test. However, the advantage of the Tukey procedure declines as the *FWE* decreases, the *df* increases, or *K* decreases. Furthermore, if only a subset of all possible pairwise comparisons are planned and tested, a point is reached at which the Dunn–Bonferroni procedure is more powerful. For example, if there are four groups but only four or fewer of the possible six comparisons are tested, the Dunn–Bonferroni method requires a smaller value of t for significance than does Tukey's method.

If the researcher plans to test a subset of all possible pairwise comparisons, the relative power of the Dunn–Bonferroni and Tukey methods is easily assessed by comparing the critical values of the two procedures when both are expressed as t statistics: Calculate the ratio of critical values of the Dunn–Bonferroni to the Tukey method:

$$D\text{–}B \text{ to Tukey ratio} = \frac{t_{FWE,K}}{q_{FWE,K}/\sqrt{2}} \tag{10.19}$$

where K is the number of comparisons, and the t and q statistics are the values required for significance when K comparisons are made. When the ratio is less than 1, the Dunn–Bonferroni requires a smaller critical value, and will therefore be the preferred method.

In summary, if the researcher carefully plans only those pairwise comparisons that are truly of interest, power may be gained by using methods that focus only on the planned comparisons. Equation 10.19 provides a basis for deciding between the methods.

10.7.3 When *ns* are Unequal: The Tukey–Kramer Test

In the *Royer* study, the number of students in the fifth–eighth grades varied; see Table 10.2 for the group means, variances, and *ns*. A modification of Tukey's *HSD* test suggested by Kramer (1956) applies in such situations. The standard t statistic is calculated and compared with $q_{FWE,a,df_e}/\sqrt{2}$. Box 10.3 illustrates the test as applied to a comparison of the fifth- and sixth-grade mean speeds.

Box 10.3 An Example of the Tukey–Kramer Test When *ns* Are Not Equal

1. Find the critical value of q. In the example of the *Royer* speed data, $a = 4$ and the error *df* are 86. If $FWE = .05$, the critical q value is approximately 3.71.[a]
2. Obtain the critical value of t: $t = q_{FWE,a,df_e} / \sqrt{2} = 3.710/1.414 = 2.623$.
3. To test the difference between the fifth- and sixth-grade mean speeds, calculate the usual t statistic presented in Chapter 6 (with s_{pooled} replaced by $MS_{S/A}$). The Tukey–Kramer t equals the mean difference ($\hat{\psi}$) value divided by the std. error ($s_{\hat{\psi}}$);

$$t = \frac{\overline{Y}_{Grade6} - \overline{Y}_{Grade5}}{\sqrt{MS_{S/Grade}\left(\frac{1}{n_6} + \frac{1}{n_5}\right)}} = \frac{.560 - .350}{\sqrt{(.032)\left(\frac{1}{23} + \frac{1}{26}\right)}}$$

$$= .21/.051 = 4.101$$

which clearly exceeds 2.623.

[a] Harter, Clemm, and Guthrie (1959) provide a method of nonlinear interpolation when the exact *df* are not in the table. For example, when the $df = 86$, find $q_{.05,4,60} = 3.74$ and $q_{.05,4,120} = 3.69$ from the table, and the reciprocals of 86 (.0116), 60 (.0167), and 120 (.0083). The critical value for $df = 86$ is then given by

$$q_{.05,4,86} = 3.74 - \left(\frac{.0167 - .0116}{.0167 - .0083}\right)(3.74 - 3.69) = 3.710$$

10.7.4 When Variances Are Unequal

In Chapter 8, in our analysis of the effects of educational level upon mean depression scores of males in the *Seasons* study, we found that the variances were quite heterogeneous. Several methods have been proposed to deal with this problem. Most use Welch's t' and df' (see Section 10.3.4) but differ in the criterion against which t' is evaluated. In the Games–Howell test (1976), t' is compared with $q_{FWE,a,df'}/\sqrt{2}$. The procedure is illustrated in Box 10.4 using the *Seasons* depression data.

Box 10.4 The Games–Howell Procedure for Testing All Pairwise Comparisons When Variances Are Not Equal

1. Compute Welch's t' and df', using Equations 6.15 and 6.16. For the *HS* and *C* statistics of Table 10.3, panel *a*, we have:

$$t' = \frac{\overline{Y}_{HS} - \overline{Y}_C}{\sqrt{\frac{s_{HS}^2}{n_{HS}} + \frac{s_C^2}{n_C}}} = \frac{6.903 - 3.674}{\sqrt{\frac{34.541}{19} + \frac{5.970}{33}}} = 2.284$$

and

$$df' = \frac{\left(\frac{s_{HS}^2}{n_{HS}} + \frac{s_C^2}{n_C}\right)^2}{\frac{s_{HS}^4}{n_{HS}^2(n_{HS} - 1)} + \frac{s_C^4}{n_C^2(n_C - 1)}} = \frac{\left(\frac{34.541}{19} + \frac{5.970}{33}\right)^2}{\frac{34.541^2}{(19^2)(18)} + \frac{5.970^2}{(33^2)(32)}} \approx 22$$

2. Obtain the critical value of t from Appendix Table C.9. For our example with 22 *df*, interpolate in Appendix Table C.9 between $df = 20$ and $df = 24$, with $a = 4$, $FWE = .05$. The critical q value is approximately 3.93. Then

$$t_{.05,4,22} = 3.93 / \sqrt{2} = 2.779$$

3. Because $2.284 < 2.779$, we cannot reject H_0: $\mu_5 = \mu_6$. In similar fashion, values of t' and df' can be calculated for each of the remaining five pairwise comparisons. Note that the critical value of t must be recalculated for each test because the df' are likely to change for each comparison.

The Games–Howell method generally does a good job of controlling *FWE* with reasonable power. However, there are some circumstances under which *FWE* may be a bit inflated with Games–Howell. If the variances are fairly homogeneous and the group sizes are less than 50, the *FWE* for the Games–Howell method may sometimes be as high as .07 when the nominal probability is .05 (Dunnett, 1980; Games, Keselman, & Rogan, 1981). Even when the variances are not homogeneous, the *FWE* may be inflated with *ns* less than 6. However, the *FWE* is close to the nominal level under most other conditions. Furthermore, the test is more powerful than any of the several competitors that have been proposed and has narrower confidence intervals for each comparison. If the researcher is concerned about the possible inflation of the *FWE*, Dunnett's T3 test (Dunnett, 1980) appears to be the most powerful of several alternatives that maintain the *FWE* at less than or equal to the nominal value. The test requires tables of the Studentized maximum modulus distribution. Miller has described the procedure (1981, pp. 70–75) and has provided tables of the distribution. The test is also available in several statistical packages.

10.7.5 The Fisher–Hayter Test

The tests considered so far do not require a preliminary test of the omnibus null hypothesis; they control the *FWE* at or below its nominal level. In fact, requiring a significant omnibus *F* prior to these tests is likely to result in a loss of power. However, there are procedures for controlling the *FWE* that include an omnibus *F* test as a first stage; pairwise comparisons are conducted only if the initial *F* test is significant. In *Fisher's* (1935) *LSD* (least significant difference) procedure, the pairwise comparisons are tested by the usual *t* test at the .05 level in a second stage. However, because the *LSD* test has been shown to have an inflated *FWE* for $a > 3$, Hayter (1986) modified the test. The resulting Fisher–Hayter test maintains the *FWE* at or below its nominal level. In this test, a significant *F* test in the first stage is followed by tests of all pairwise comparisons using a standard *t* test, but each test is evaluated against the criterion $q_{FWE,\,a-1,\,df_e}/\sqrt{2}$. Note that in entering Appendix Table C.9, the column corresponding to $a-1$ means provides the critical value.

Note that the Fisher–Hayter test is a sequential procedure involving two steps. As with other sequential testing methods, power is gained relative to simultaneous tests but at the loss of the ability to construct simultaneous confidence intervals. Therefore, the choice between the Fisher–Hayter test and the Tukey (or Tukey–Kramer) test depends on whether such confidence intervals are desired. Another consideration is whether variances are assumed to be homogeneous. The Fisher–Hayter test was derived under that assumption and therefore a test such as the Games–Howell or the Dunnett T3 should be used if homogeneity of variance is in doubt.

10.7.6 Pairwise Comparisons: Summing Up

The rather detailed set of recommendations for controlling *FWE* over families of pairwise comparisons results from an attempt to satisfy two criteria. The first is that we want a method that adequately controls the Type 1 error rate over the family. We have seen in previous chapters that two factors that often affect Type 1 error rates are heterogeneous variances and unequal *n* across conditions. We again find that the same two factors require adjustments of our procedures for controlling familywise error rates.

Given that we can identify alternative procedures that satisfactorily control Type 1 error rates, our second consideration is to choose the procedure that has the most power. Seaman et al. (1991) simulated tests of all pairwise comparisons under conditions of equal *n*s and equal variances. Over the conditions examined in their study, Seaman et al. found that the Fisher–Hayter method had a power advantage over the Tukey *HSD* method that varied between 2% and 9%; Tukey, in turn, had a power advantage of 2–3% over the Dunn–Bonferroni method. The other distinction among procedures that is relevant to power concerns is the distinction between simultaneous and sequential methods. Sequential tests have more power than simultaneous tests, although confidence intervals are not available when sequential methods are used.

Box 10.5 summarizes our recommendations with respect to the control of *FWE* over families of pairwise comparisons.

Box 10.5 Recommendations for Controlling *FWE* on Families of Pairwise Comparisons

1. If the researcher wants to construct confidence intervals to estimate the differences between group means:
   a. If the variances are homogeneous and *n*s are equal, use Tukey *HSD*.
   b. If the variances are unequal, use Games–Howell (or Dunnett T3 if the *n*s are fewer than 6).
   c. If the *n*s are unequal, use Tukey–Kramer.
   d. If only a subset of all pairwise comparisons are planned, use Equation 10.19 to determine whether Tukey HSD or Dunn–Bonferroni will have more power.
2. If the researcher only wants to conduct hypothesis tests:
   a. If the variances are homogeneous, use Fisher–Hayter.
   b. If the variances are unequal, use Games–Howell (or Dunnett T3 if the *n*s are fewer than 6).

We have excluded from our discussion of pairwise comparison procedures that are sometimes used; namely, Fisher's *LSD* test (1935), the Student–Newman–Keuls test (Keuls, 1952; Newman, 1939), and Duncan's multiple range test (1955). We advise against the use of these procedures because they yield *FWE*s that often are considerably in excess of the nominal value. We have also excluded several procedures that maintain the *FWE* at or below its nominal level and have slightly more power than the Fisher–Hayter method under some combinations of number of groups and group size (e.g., Peritz, 1970; Ramsey, 1978, 1981; Shaffer, 1979, 1986; Welsch, 1977). On the basis of various sampling studies, the very slight power advantage of these methods (usually 1% or 2%) does not warrant the added complexity they usually entail. Descriptions of these methods, together with results of sampling experiments, may be found in the article by Seaman et al. (1991); multiple comparison procedures are also reviewed by Zwick (1993), Shaffer (1995), and Toothaker (1993).

Finally, we note that some or all of the pairwise comparison procedures we have considered, as well as the Dunn–Bonferroni and Dunn–Šidák tests, are available in various statistical software packages. For example, SPSS can perform 12 different tests of pairwise comparisons (select the *Post hoc* option in the *Compare Means/One-Way ANOVA* or in the *General Linear Model/Univariate* menu). Although it is tempting to run several of these tests, we urge researchers to select one procedure in advance, and base conclusions on the results of that test.

10.8 COMPARING *a* − 1 TREATMENT MEANS WITH A CONTROL: DUNNETT'S TEST

Dunnett (1955, 1964) proposed a test for studies in which the researcher plans to contrast each of several treatments with a control. If these are the only comparisons of interest, methods that control the *FWE* for a family consisting of *all* pairwise comparisons will be overly conservative; power will be lost and simultaneous confidence intervals will be wider than necessary. The Dunn–Bonferroni procedure with $K = a - 1$ will be an improvement but will still offer less power and wider intervals than the Dunnett test.

Assuming that the group sizes are equal and that variances are homogeneous, the test is quite simple. Box 10.6 illustrates the procedure using the data from the memory study. However, if

Box 10.6 Dunnett's Test Comparing the Experimental Means with the Control Mean in the Memory Study

1. Compute the usual t statistic comparing the control with each expermental group; e.g., to compare the control and loci means

$$t = \frac{\bar{Y}_{loci} - \bar{Y}_C}{\sqrt{MS_{S/A}\left(\frac{2}{n}\right)}} = \frac{12.1 - 6.5}{\sqrt{17.944\left(\frac{2}{10}\right)}} = 2.96$$

   and for the comparison of the control group mean with the rhyme and image means, $t = 2.11$ and 2.22, respectively.
2. Evaluate the three t statistics against the critical value of $d_{FWE,a,df}$ in Appendix Table C.8, where a is the number of means including the control and df is the number of degrees of freedom associated with the ANOVA error term. In the present example, FWE (two-tailed) $= .05$, $a = 4$, and the error $df = 36$; the critical value is 2.48. Only the control and loci means differ significantly.
3. The confidence intervals have the same form as in the two-independent-group examples of Chapter 6; the bounds are

$$(\bar{Y}_j - \bar{Y}_C) \pm s_{\hat{\psi}} d_{FWE,a,df}$$

   For example, for comparison of the loci mean with the control mean, the bounds are

$$(12.1 - 6.5) \pm \sqrt{17.944\left(\frac{2}{10}\right)} \times 2.48 = .90,\ 10.30$$

the group sizes are not equal, replace $2/n$ in the equation for t by $1/n_j + 1/n_C$, and use the Dunn–Bonferroni procedure with $K = a - 1$. If any of the a group variances differ, use Welch's t' and again use the Dunn–Bonferroni procedure.

10.9 CONTROLLING THE FAMILYWISE ERROR RATE FOR POST HOC CONTRASTS

Sometimes observed patterns in the data suggest the presence of effects that had not been anticipated and that are not adequately captured by the set of all possible pairwise comparisons. When the corresponding null hypotheses are tested to determine whether these effects are significant, we should be quite conservative in evaluating the result. In testing contrasts "after the fact" we are, in effect, investigating the family of all possible outcomes. Therefore, the methods we present are quite conservative because they control for the probability of at least one Type 1 error in a very large set of possible contrasts.

10.9.1 Scheffé's Method

Assuming that the populations are normally distributed and have equal variances, Scheffé's (1959) method maintains the FWE at its nominal level when the family consists of all possible contrasts associated with a source of variance. Using the fifth–eighth-grade multiplication speeds in the study by Royer et al. (1999) as an example, assume that we had not anticipated the pattern of means in Table 10.2. After viewing the data, we observe that the sixth–eighth grades had very similar means, each higher than the fifth-grade mean. We might wish to test whether the mean of the fifth-grade response times differs significantly from that of the three combined sixth–eighth-grade times. Box 10.7 describes the Scheffé procedure, and illustrates its application to the contrast of the fifth-grade mean with the average of the other three means.

Box 10.7 Scheffé's Method to Test H_0: $(1/3)(\mu_6 + \mu_7 + \mu_8) - \mu_5 = 0$ (*Royer* speed data)

1. Calculate the t statistic to test the contrast of interest (see Equation 10.5).
2. Compare the computed value of t with $S = \pm\sqrt{df_1 \cdot F_{FWE,df_1,df_2}}$ where df_1 and df_2 are the numerator and denominator degrees of freedom.
3. For the arithmetic experiment, $df_1 = 3$ and $df_2 = 36$; if $FWE = .05$, the critical F is approximately (from Appendix Table C.5) $F_{FWE,df_1,df_2} = 2.88$. Therefore,

$$S = \pm\sqrt{(3)(2.88)} = 2.94$$

4. Reject the null hypothesis if $t > S$ or $t < -S$. To test the null hypothesis, $t = 5.23$. Because $5.21 > 2.94$, we reject H_0.
5. The formula for the confidence interval bounds is

$$\hat{\psi} \pm s_{\hat{\psi}} \sqrt{df_1 \cdot F_{FWE,df_1,df_2}}$$

   where $\hat{\psi} = .680$ and $s_{\hat{\psi}} = \sqrt{MS_{S/A}(\Sigma w_j^2 / n_j)} = .130$. Therefore, the bounds on $(\mu_6 + \mu_7 + \mu_8) - 3\,\mu_5$ are $.680 \pm (.130)(2.94) = .298,\ 1.062$.
6. To return to the original scale, these bounds must be divided by 3; the bounds on $(1/3)(\mu_6 + \mu_7 + \mu_8) - \mu_5$ are .099 and .354.

It is instructive to compare the confidence interval presented in Box 10.7 with the results we would have obtained if our contrast had been planned. Assume that the contrast was one of three planned for the experiment. In that case, we could have used the Dunn–Bonferroni method to compute the confidence interval. In contrast to the interval limits in Box 10.7, the Dunn–Bonferroni limits are

$$\hat{\psi} \pm t_{FWE/K} s_{\hat{\psi}}$$

The contrast and its standard error are .680 and .130 (see Box 10.7), and the t required for significance at the $.05/3 = .0167$ level (two-tailed) is 2.51. Substituting these values (and dividing the resulting limits by 3 to return to the original scale), we find the Dunn–Bonferroni limits to be .118 and .335. The Dunn–Bonferroni interval is narrower than the Scheffé interval in Box 10.7, revealing the price we pay in precision of estimation and power when contrasts are not planned. Whenever possible, it is a good strategy to plan all contrasts that might conceivably be of interest, and then use the Dunn–Bonferroni or Fisher–Hayter method. Although the power of these methods decreases as the number of planned contrasts increases, a rather large number of comparisons must be planned before the Scheffé criterion requires a smaller value of t for significance (see Perlmutter & Myers, 1973, for a more detailed comparison of the Dunn–Bonferroni and the Scheffé methods).

Experimenters who have used both the standard ANOVA tests and the Scheffé procedure have sometimes been surprised to find that the omnibus null hypothesis is rejected by the ANOVA test but that no contrasts are significant by the Scheffé criterion. The source of this apparent contradiction is that the overall F test has exactly the same power as the *maximum possible contrast* tested by the Scheffé procedure. That contrast may be of little interest, so it may not have been tested. It could be something like $(11/37)\mu_1 + (26/37)\mu_2 - (17/45)\mu_3 - (28/45)\mu_4$. In summary, although rejection of the omnibus null hypothesis indicates that at least one contrast is significant by the Scheffé criterion, there is no guarantee that any obvious or interesting contrast will be significant.

As with all the tests we have so far considered (except the Fisher–Hayter), there is no logical necessity that the Scheffé tests of contrasts be preceded by a significant omnibus F. On the other hand, if the omnibus F test is not significant, no contrast will be significant. Thus, there is little point in expending energy on a series of post hoc Scheffé tests unless first determining that the F test is significant.

10.9.2 The Brown–Forsythe Method When Variances Are Not Equal

Brown and Forsythe (1974b) proposed that Welch's t' and df' (Box 10.3) be used with a criterion similar to Scheffé's S when the test is post hoc and the assumption of homogeneity of variance is questionable. The only difference is that the critical value of S against which t' is evaluated is based on df' (see Equation 10.10).

10.10 CONTROLLING THE FAMILYWISE ERROR RATE IN MULTI-FACTOR DESIGNS

We have been considering the control of *FWE* in the context of examples taken from one-factor designs. Although the calculations and methods for control of error rates are the same in multi-factor designs, there are several additional issues. Consider a two-factor design with four levels of B (e.g., type of drug) and two levels of A (e.g., age). We may wish to compare the B marginal means to determine the relative efficacy of the different drugs. We might use the Tukey *HSD* procedure to control *FWE*. We may also wish to compare the means of the B conditions within each level of A (that is, the simple effects of drug at each age). Is each level of A a family with the *FWE* set at .05? Or should *FWE* be set at .025 at each level of A so that the *FWE* is .05 for the complete set of tests? And should the comparisons of marginal means and of simple effects be considered separate families, or one family? Suppose we are also concerned with testing interaction effects? Is this still another family of tests? Or should all the tests be considered a single family, thus controlling Type 1 error rates simultaneously for all hypotheses tested in the experiment, but sacrificing power? There are no generally agreed-upon answers to such questions. We will make some recommendations; however, depending on their designs and the questions they wish to address, investigators may decide on different approaches than the one we take in this section. Whatever the approach to controlling *FWE*, it is important that any report of research be clear about just how the *FWE* was controlled so that readers may perform their own evaluation of the significance of results.

In the remainder of this section, we illustrate some common tests of contrasts, and our recommendations for controlling the *FWE*, using a simple 2 × 4 set of means, each based on six scores. Table 10.5 contains the cell means and the A and B marginal means, together with the ANOVA summary.

Table 10.5 Cell means (*a*) and ANOVA (*b*)

(*a*) Cell means

	Drug type				
Age group	B_1	B_2	B_3	B_4	Mean
A_1	12	6	5	15	9.5
A_2	16	2	9	3	7.5
Mean	14	4	7	9	

(*b*) ANOVA ($n = 6$)

Source	*df*	*SS*	*MS*	*F*	*p*
Age group (A)	1	48	48	1.280	.265
Drug (B)	3	638	212.67	5.653	.003
AB	3	528	176	4.693	.007
S/AB	40	1,500	37.5		

10.10.1 Testing Hypotheses About Marginal Means

The results of the analysis of variance in Table 10.5 reveal that the drug type (B) has significant effects, and the significant interaction suggests that the sizes of these effects are different in the two age groups (A). Let us consider the effects of B first. Most likely, we would wish to compare pairs of the four drug means. Tukey's *HSD* method provides a way to control the familywise error rate for the six pairwise tests we will perform. Setting the *FWE* at .05 and turning to Appendix Table C.9, we find that the critical value of q, the Studentized range statistic, is 3.79 when $a = 4$ and the error $df = 40$. The standard error of the mean is $SEM = \sqrt{MS_{S/AB} / n} = 1.768$. Note that $n = 12$, which is the number of scores contributing to each mean at each level of B. As in Box 10.2, we multiply the critical value of q by the *SEM*, yielding a critical difference of 6.700. Only the difference between the marginal B_1 and B_2 means exceed this value; therefore, controlling the *FWE* at .05, the only significance pairwise difference is between these two drugs.

Sometimes the researcher may have an *a priori* hypothesis that specified that only certain comparisons, whether pairwise or more complex, would be significant. In that case, the Dunn–Bonferroni method should be followed. The per comparison error rate would be the desired familywise error rate divided by K, the number of tests; this would be the critical p-value for our significance tests.

Another possible scenario is that the researcher observes the means in Table 10.5 and decides at that point that the difference between the B_1 and B_2 means is large enough to warrant further investigation. Or, after viewing the means, the researcher decides that the B_1 mean is different enough from the other means that it should be tested against the average of the other three means. Such post hoc contrasts should be tested by the Scheffé procedure described in Box 10.7.

Two points should be noted about the developments so far. First, we have identified a family as a set of comparisons related to a single source of variance. If we had several levels of A, we also would have controlled the *FWE* at .05 for all comparisons of those marginal means. The second point is that we have assumed homogeneity of variance. As we discussed earlier in this chapter, modifications of the usual tests are indicated when this assumption is not met. In particular, the Tukey *HSD* method is not appropriate because the various means have different standard errors. In such a case, even if all pairwise comparisons are of interest, the Bonferroni procedure should be applied with the standard error of the difference between means based on the average within-cell variance of the cells involved in each comparison.

In sum, contrasts to analyze main effects within multi-factor designs follow the same recommendations and procedures presented earlier in the context of the one-factor design.

10.10.2 Testing Hypotheses About Simple Effects

Having found a significant difference between the marginal B_1 and B_2 means, we may ask whether this difference is significant in either or both age groups. Also, although no other differences among the marginal means were significant, it is possible that there are other significant differences in one of the two age groups. Therefore, a reasonable next step is to compare the drugs at each level of *A*. The issue is what familywise error rate will be acceptable. We view the set of all contrasts of simple effects as a single family and therefore recommend that the *FWE* at each level of *A* be .05 divided by the number of levels of *A*. In the example of Table 10.5, pairwise comparisons within each of the two age groups would be tested with the critical value of *q* at the .025 level. Although Appendix Table C.9 contains only .01, .05, and .10 values for the Studentized range, most software that performs the Tukey *HSD* test will allow entry of any criterion for significance; values are also available in the original Harter et al. (1959) technical report from which our table was adapted. The required value of *q* is 4.197 when $a = 4$ and the error $df = 40$.

It may seem that a first step would be to perform an omnibus *F* test at each level of *A*. However, the Studentized range statistic when applied to the largest difference between means is a test of the omnibus null hypothesis that all means are equal.[5] Therefore, we proceed directly to pairwise comparisons at each level of *A*. Assuming that the eight population variances are homogeneous, the critical difference is the standard error times the critical value of *q*; i.e., $d_{crit} = 2.5 \times 4.197$, or 10.493. At A_1, the difference between the B_4 and B_2 means, and that between the B_4 and B_3 means are the two largest, but neither exceeds the critical value. Accordingly, we cannot reject the hypothesis that the *B* population means at A_1 are equal. At A_2, however, there are significant differences; the B_1 mean differs significantly from the B_2 and the B_4 means. Our analysis of the simple effects of the drugs has provided important information. Although we cannot be sure that other differences among the drug effects do not exist (because we cannot accept the null hypothesis), we have found that in the second age group drug B_1 is clearly superior to drugs B_2 and B_4.

We have assumed that all pairwise comparisons within each level of a second factor are to be performed. There may be circumstances in which only some of the possible comparisons within each level are of interest. For example, assume that we have *a priori* hypotheses about three contrasts at A_1 and two more contrasts at A_2. In that case, the Dunn–Bonferroni procedure with $K = 5$ is appropriate; each test would be performed with $\alpha = .01$. On the other hand, assume that the same contrasts appear to be of interest only after viewing the cell means. In that case, the Scheffé procedure, described in Box 10.7, is appropriate with $df_1 = (a - 1)(b - 1)$.

[5] See Myers, 1979, for a discussion of the relative power of the *F* and Studentized range tests of the omnibus null hypothesis.

10.10.3 The Relation Between Tests of Interactions and Tests of Simple Effects

Contrasts of simple effects are of interest in their own right and, when the error rate is properly controlled, can be helpful in drawing more precise inferences about the effects of our variables. However, a common misconception is that such tests are performed only to help us understand the causes of a significant interaction. They may sometimes do so, particularly when differences among means are significant at one level of a factor but not at others, as in our analysis of the means in Table 10.5. In many cases, however, the pattern of results is inconsistent with what we would expect if our means were population means. We may have a significant interaction and fail to find any significant difference between means of one variable at any level of the second. Or the interaction may fail to be significant, but simple effects of one variable may differ significantly. The tests of interaction and of simple effects differ in power both because of differences in the criteria for significance and because of differences in the cell frequencies and variances involved.

A common result that sometimes puzzles researchers is illustrated by the following cell means:

	B_1	B_2
A_1	15	5
A_2	8	6

Assume nine scores in a cell and an average cell variance of 100. Then, the *t* test of interaction is

$$t_{interaction} = [(15 - 5) - (8 - 6)] / (10\sqrt{4/9}) = 1.20$$

The interaction is not significant. However, if we compare the simple effects of *B* at A_1 we have

$$t_{B/A_1} = (15 - 5) / (10\sqrt{2/9}) = 2.12$$

which is significant at the .021 level. The test at A_1 does not yield a significant result. Even controlling the *FWE* at .025 for each test comparing simple effects, we conclude that there is a difference between the *B* means at A_1. We now have an apparent contradiction: The test of the interaction does not provide evidence of an interaction in the population. However, the significant result of the test at A_1, coupled with a nonsignificant result of the test at A_2 (that *t* was .42), suggests that there is an interaction with *B* having effects at A_1 but not at A_2. The reason for this inequality is that the standard error of the interaction involves the variance of differences among four means, whereas the standard error for the simple effect involves the variance of only two means. Therefore, it is possible to have a pattern of test results for simple effects that is not consistent with the result of the test of interaction because the power of the test of the simple effects is greater than the power of the test of the interaction.

10.10.4 Testing Hypotheses About Interaction

Various proposals have been made about the proper follow-up analyses to perform to understand a significant interaction (Games, 1973; Marascuilo & Levin, 1972, 1973; Tukey, 1991). The most direct approach is to test embedded 2 × 2 interactions. For example, in Table 10.5, we might test the interaction involving the two age groups and the B_1 and B_4 drugs. This interaction effect size is

$(12 + 3) - (15 + 16) = -16$. Dividing this by its standard error, $\sqrt{MS_{S/AB}(4/n)}$, we have the t statistic or, squaring, we have the F. If this test had been planned, the alpha level would be set at .05. However, if such analyses are performed, they are typically post hoc. In that case, we have two options for controlling the familywise error rate. Assuming homogeneous variances, we can apply Scheffé's method; we calculate $S = \sqrt{(a-1)(b-1)F_{.05,(a-1)(b-1),ab(n-1)}}$. In our example, $(a-1)(b-1) = 3$ and $ab(n-1) = 40$. The F required for significance at the .05 level with 3 and 40 df is 2.84. Therefore, $S = 2.92$. An alternative is to view the set of six possible 2 × 2 interaction contrasts as a family and use the Bonferroni criterion. With the FWE set at .05, the alpha level for each t test is .05/6, or .0083. Assuming a two-tailed test, the critical value of $t = 2.78$, slightly smaller than the critical value of S. As in this example, the Bonferroni criterion will often be slightly more powerful though its advantage will be lost as the number of interaction contrasts increases. As in our example, when variances are homogeneous, the choice between methods rests on a comparison of the critical values (see Perlmutter & Myers, 1979). If variances are heterogeneous, the Bonferroni method should be used, with the error term and error df based on the contrast tested.

10.10.5 Using Software to Perform Further Analyses

In most software packages, simple effects can readily be tested by splitting the file. For example, using SPSS to test all pairwise comparisons among drugs for each age group, we would select the *Data* menu, and then indicate that the file is to be split by levels of A. A univariate analysis would then be performed on each of the a sets of means after selecting the Tukey HSD test from the available post hoc options. The *select cases* option available in most packages will enable selection of specified levels of factors, permitting tests of embedded 2 × 2 interactions. However, there are two caveats to keep in mind when splitting files or using *select cases* to conduct contrasts on subsets of observations in a data file. First, the error terms for the tests will be based just on the subset of observations selected for analysis. If variances are homogeneous in the experiment and the researcher wishes to take advantage of the increased number of df that result from pooling the variances across all conditions (i.e., using $MS_{S/AB}$ in the error term of each contrast), the researcher will need to recompute the denominator of the contrast. The second caveat is that the reported p-values are unlikely to reflect control of the familywise error rate. In general, the researcher must ensure that the appropriate criterion for significance has been applied, either through options in the software or, when these are lacking, by comparing the test statistic against the appropriate criterion.

10.11 SUMMARY

This chapter developed the following points:

- Contrasts are specific comparisons on a set of means that allow researchers to pose detailed questions of a data set.
- Procedures for conducting hypothesis tests and constructing confidence intervals on contrasts are straightforward extensions of the t test procedures covered in Chapter 6.
- In an experimental design of any complexity, there are many possible contrasts that might be of interest to a researcher. The probability that at least one Type 1 error will occur in a set, or family, of contrasts increases as a function of the size of the family. It is therefore important to control the probability of a Type 1 error across a family (i.e., FWE).

- There are many different kinds of families of contrasts of means. These include the family of comparisons planned prior to data collection; the family of all possible comparisons of members of pairs of group means; the family of all comparisons of experimental group means with a control mean; and the family of post hoc contrasts determined on the basis of viewing the data.
- Different methods have been developed for controlling the FWE, depending on the kind of family, and on whether confidence intervals are desired. Most of these methods involve the usual t statistic, or its close relative, the Studentized range statistic, or—when heterogeneity of variance is suspected, Welch's t'. The major difference among the methods is the criterion employed for judging a contrast of means to be significant.

Table 10.6 provides a summary of much of what has been presented in this chapter. This summary integrates several considerations affecting the choice of procedures for controlling Type 1 error. One reason for the many different procedures in Table 10.6 is that the control of Type 1 errors is influenced by considerations such as whether the assumption of homogeneity of variance has been met, and whether ns are equal across conditions. It is paramount that a given procedure controls Type 1 errors at close to the nominal level. Given that this criterion is met, power considerations are the second major reason for the many procedures presented in Table 10.6. Given two procedures that adequately control Type 1 error rates, we prefer the method that results in more powerful tests. Sequential methods provide more powerful tests, so they are preferred over simultaneous methods when hypothesis tests are conducted. However, sequential methods are not applicable to the construction of confidence intervals because there is no ordering within a set of confidence intervals. Finally, we emphasize again that contrasts should be planned whenever possible because a set of planned contrasts is almost always smaller than a set of unplanned contrasts. Thus, tests of planned contrasts generally have more power than tests conducted on other kinds of families. But perhaps the more important benefit of planning contrasts is that the careful thought that is required to specify the key research questions will probably lead to research designs that are more closely focused on those questions.

Table 10.6 Recommended procedures for controlling *FWE*

Family type	Simultaneous methods[a]		Sequential methods	
	Equal variances	Unequal variances	Equal variances	Unequal variances
Planned	Dunn–Bonferroni	Dunn–Bonferroni using Welch's t'	Hochberg	Hochberg using Welch's t'
All pairwise	Tukey HSD (equal n) or Tukey–Kramer (unequal n)[b]	Games–Howell or Dunnett T3	Fisher–Hayter	
Exptl vs control	Dunnett (equal n) or Dunn–Bonferroni (unequal n)	Dunn–Bonferroni using Welch's t'		
Post hoc	Scheffé	Scheffé using Welch's t'		

[a] Only the simultaneous methods allow the construction of simultaneous confidence intervals.

[b] Assuming K pairwise tests, the Bonferroni method will be more powerful than the Tukey under some conditions; Equation 10.19 provides the basis for the choice.