A Series of Books in Psychology

Data Analysis for Research Designs

* Analysis of Variance and Multiple Regression/Correlation Approaches

(* Not covered by Erick).

GEOFFREY KEPPEL
SHELDON ZEDECK
University of California at Berkeley

W. H. FREEMAN AND COMPANY
NEW YORK

1989

The previous chapters laid the foundation for the basic statistical concepts in research and statistics. Now it is time to relate specific analytical statistics to research design. This chapter deals with the analysis of the simplest experimental design, one in which two groups of individuals are treated differently and the effects of the differential treatment are assessed. We will show how the results of this simple experimental design are analyzed by means of the analysis of variance (ANOVA). In Chap. 7 we will consider the corresponding analysis with MRC procedures.

6.1 SUBDIVIDING THE TOTAL SUM OF SQUARES

Suppose we examined the data from an experiment in which two groups received different treatments but we did not take this fact into consideration in our analysis. Although we could calculate a mean and a standard deviation to describe the *total* data set, they would not be particularly useful, since this summary would obscure any effects that the two different treatments might have. Of course, these effects are precisely what we are most interested in. The analysis of variance does start, however, with a sum of squares based on the total data set. This *SS* is subdivided (or partitioned) into a number of separate parts, each of which provides different information useful for the statistical analysis of the experiment.

Notation and Labels

Let us consider the outcome of a hypothetical experiment in which fifth-grade students are introduced to a set of vocabulary words presented in the context either of a lecture on physical science or of one on social science. A vocabulary test, with 60 words, follows the lecture. The purpose of the experiment is to determine which of the two lectures produces better performance on the vocabulary test. The researcher begins with a total of 24 subjects and randomly assigns them so that each will fall into one of the two lecture conditions; equal numbers of subjects (12) are assigned to each group. We will use s to represent the treatment group sample size and N to represent the total number of subjects in the experiment. In this example, then, group sample size is $s = 12$, and $N = 12 + 12 = 24$. (We will assume equal sample sizes for most of the experimental designs we consider in this book; Chap. 24 deals with unequal sample sizes.)

We will call the independent variable **factor** A. The specific treatment conditions (or "treatments," for short) will be referred to as the **levels** of the independent variable, symbolized as *level* a_1 and *level* a_2. A lowercase a without a subscript designates the number of levels constituting factor A. In this example, there are $a = 2$ levels: level a_1 (physical science) and level a_2 (social science). The total number of subjects N is specified by multiplying the number of treatment levels or conditions by the number of subjects per level—that is, $N = (a)(s) = (2)(12) = 24$.

The results of the experiment are presented in Table 6-1. The individual scores Y on the vocabulary test can range from 0 to 60. You will note that the notational system has been expanded in order to specify precisely the treatment membership of any given score. The symbol for the individual observations remains the same (Y), with subscripts added so that we can refer to specific scores. Generally, we will use subscripts only when necessary to avoid confusion.

The sums or totals of the scores for the two different treatment groups are specified by the notation A_1 and A_2; thus, A_1 is the sum of the Y scores for the 12 children assigned to level a_1 (physical science) and A_2 is the corresponding sum for the 12 children assigned to level a_2 (social science). The grand total of the scores (or sum of all the scores) is symbolized by T. Applying this notation to the present example, we have

$$A_1 = 53 + 49 + \cdots + 32 + 27 = 480$$
$$A_2 = 47 + 42 + \cdots + 11 + 6 = 312$$
$$T = A_1 + A_2 = 480 + 312 = 792$$

Table 6-1
Numerical Example

	Physical Science (a_1)	Social Science (a_2)
	$Y_1 = 53$	$Y_{13} = 47$
	$Y_2 = 49$	$Y_{14} = 42$
	$Y_3 = 47$	$Y_{15} = 39$
	$Y_4 = 42$	$Y_{16} = 37$
	$Y_5 = 51$	$Y_{17} = 42$
	$Y_6 = 34$	$Y_{18} = 33$
	$Y_7 = 44$	$Y_{19} = 13$
	$Y_8 = 48$	$Y_{20} = 16$
	$Y_9 = 35$	$Y_{21} = 16$
	$Y_{10} = 18$	$Y_{22} = 10$
	$Y_{11} = 32$	$Y_{23} = 11$
	$Y_{12} = 27$	$Y_{24} = 6$
Sum:	$A_1 = 480$	$A_2 = 312$
Mean:	$\bar{Y}_{A_1} = 40.00$	$\bar{Y}_{A_2} = 26.00$

The symbols for the two treatment means are $\bar{Y}_{A_1}$ and $\bar{Y}_{A_2}$. Each is based on $s = 12$ observations in this example. The grand mean of all the scores is designated $\bar{Y}_T$ and is based on $N = (a)(s) = (2)(12) = 24$ observations. For these data,

$$\bar{Y}_{A_1} = \frac{A_1}{s} = \frac{480}{12} = 40.00$$

$$\bar{Y}_{A_2} = \frac{A_2}{s} = \frac{312}{12} = 26.00$$

$$\bar{Y}_T = \frac{T}{(a)(s)} = \frac{792}{24} = 33.00$$

The Deviations

We have plotted the data from Table 6-1 in Fig. 6-1, using filled rectangles for the scores from level a_1 and unfilled rectangles for the scores from level a_2. Consider the final score listed for a_2 in Table 6-1, namely, $Y_{24} = 6$. The deviation of this score from the grand mean, $Y_{24} - \bar{Y}_T$, is indicated at the bottom of Fig. 6-1. We will call this deviation the **total deviation.** It is readily apparent that the total deviation is made up of two parts, namely, the deviation of the individual score from its group mean, or $Y_{24} - \bar{Y}_{A_2}$, and the deviation of the group mean from the grand mean, $\bar{Y}_{A_2} - \bar{Y}_T$. We will call these deviations the **within-group deviation** and the **between-groups deviation**, respectively. Thus, we see that a single score can be viewed in terms of how it differs from the total sample, how it differs from other scores in its group, and, indirectly, how its group differs from the total sample. The deviations of Y_{24} are expressed in numbers as follows:

$$Y_{24} - \bar{Y}_T \quad = \quad (Y_{24} - \bar{Y}_{A_2}) \quad + \quad (\bar{Y}_{A_2} - \bar{Y}_T)$$

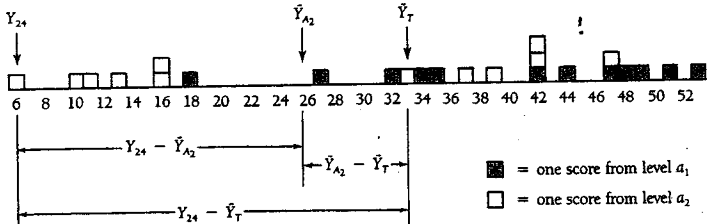


Figure 6-1 A systematic arrangement of scores from Table 6-1. The components of deviation for a single score (Y_{24}) from the grand mean ($\bar{Y}_T$) are shown beneath the baseline.

Total deviation = within-group deviation + between-groups deviation

$$\begin{array}{ccccc} 6 - 33 & = & (6 - 26) & + & (26 - 33) \\ -27 & = & (-20) & + & (-7) \\ -27 & = & -27 & & \end{array}$$

These two component deviations provide useful information concerning the outcome of an experiment. The within-group deviations, for example, represent the variability of subjects treated alike, that is, how subjects still differ even though they are in the same treatment condition. Because subjects in a particular treatment group are all given the same treatment—either a lecture on physical science or one on social science—any remaining variability in Y scores among the subjects within a group must be due to factors other than the differences between the two treatments. We will refer to any such differences as *uncontrolled variability*. An alternative way of describing these same differences is to characterize them as reflecting variability on the dependent variable Y that is *not attributable* to or "explained" by the manipulation of the treatment.

The between-groups deviations, on the other hand, represent that part of the total deviation that is associated with the two treatment conditions. You should realize that differences between groups will nearly always be present even if the independent variable is *completely ineffective*. This is due in large part to the fact that subjects are assigned *randomly* to the different conditions, creating differences between the group means that result entirely from chance. There is no satisfactory way of avoiding this problem. For this reason, then, the between-groups deviation is assumed to reflect the *joint presence* of two factors, chance differences as well as the possible differential effects of the treatments themselves.

6.2 SUMS OF SQUARES: COMPUTATIONAL FORMULAS

The partitioning presented in Sec. 6.1 can be applied to all the scores in an experiment. The three sets of deviations, when squared and summed, will produce three corresponding sums of squares, namely,

SS_T: the sum of squares based on the total deviations (that is, $Y - \bar{Y}_T$), where the subscript T refers to the total variability

$SS_{S/A}$: the sum of squares based on the within-group deviations (that is, $Y - \bar{Y}_A$), where the subscript S/A stands for the variability of subjects (S) within each of the levels of factor A (A)

SS_A: the sum of squares based on the between-groups deviations $(\bar{Y}_A - \bar{Y}_T)$, where the subscript A refers to variation associated with the independent variable, factor A

In addition, you should note that the two component sums of squares $SS_{S/A}$ and SS_A combine to equal the total sum of squares SS_T. That is,

$$SS_T = SS_{S/A} + SS_A \qquad (6\text{-}1)$$

These three sums of squares are the basic components of ANOVA. When you are computing sums of squares with a calculator, you will find it much easier to use computational formulas than to deal with the three sets of deviations for each subject. We will consider the appropriate computational formulas next.

Basic Ratios

Computational formulas for any sum of squares can be expressed in terms of what we will call **basic ratios.** Because all basic ratios involve the same set of simple arithmetic operations, you should take note of these consistent operations, since they will be found in all standard analyses of variance. The present design requires three such ratios, one based on the Y scores, another based on the A treatment sums, and a third based on the grand sum T. Each set of terms contributes to the numerator of a different basic ratio. More specifically, all members of a given set of quantities—the Y's, the A's, and T—are first *squared* and then *summed.* (In the case of T, where there is only one quantity, just the first operation is performed.) Using the data from Table 6-1, we obtain

$$\Sigma\, Y^2 = 53^2 + 49^2 + \cdots + 11^2 + 6^2 = 31{,}136$$
$$\Sigma\, A^2 = 480^2 + 312^2 = 230{,}400 + 97{,}344 = 327{,}744$$
$$T^2 = 792^2 = 627{,}264$$

Each of these numerators is divided by a different number, which is found by applying a simple rule that involves the term appearing in the numerator:

> **Whatever the term—that is, Y, A, or T—we divide by the number of scores that contribute to that term.**

For Y this number is 1, because each Y score is based on a *single* observation; this is equivalent, of course, to not dividing at all. For A this number is s, because this is the number of scores that are summed to produce any one of the treatment sums, while for T this number is $(a)(s)$, or N, because this is the number of scores that are actually summed to produce the grand sum.

For convenience, each basic ratio is given a special symbol consisting of a pair of brackets enclosing the letter code used to designate terms in the numerator. The formulas for the three basic ratios are

$$[Y] = \Sigma\, Y^2 \qquad (6\text{-}2)$$

$$[A] = \frac{\Sigma\, A^2}{s} \qquad (6\text{-}3)$$

$$[T] = \frac{T^2}{(a)(s)} \qquad (6\text{-}4)$$

Applying these formulas to the partial answers we have already calculated, we find

$$[Y] = 31{,}136$$

$$[A] = \frac{327{,}744}{12} = 27{,}312.00$$

$$[T] = \frac{627{,}264}{(2)(12)} = 26{,}136.00$$

Sums of Squares

The three sums of squares are easily calculated by combining the basic ratios in different patterns. These patterns are specified by the deviations themselves. You will recall that the total sum of squares is based on the following deviation:

$$Y - \bar{Y}_T$$

The computational formula combines the two ratios identified by this deviation as follows:

$$SS_T = [Y] - [T] \qquad (6\text{-}5)$$

where $[Y]$ is the basic ratio based on the individual Y scores or observations and $[T]$ is the basic ratio based on the grand sum T.

The within-groups sum of squares is based on the deviation of individual observations from the relevant treatment mean, that is,

$$Y - \bar{Y}_A$$

The computational formula combines the two ratios identified by these deviations as follows:

$$SS_{S/A} = [Y] - [A] \qquad (6\text{-}6)$$

where $[Y]$ is the basic ratio based on the individual Y scores and $[A]$ is the basic ratio based on the two treatment sums.

Finally, the between-groups sum of squares is based on the deviation of the treatment means from the grand mean:

$$\bar{Y}_A - \bar{Y}_T$$

The computational formula combines the two ratios identified by these deviations as follows:

$$SS_A = [A] - [T] \qquad (6\text{-}7)$$

where [A] is the basic ratio based on the treatment sums and [T] is the basic ratio based on the grand sum.

We will now calculate these sums of squares by substituting in these three formulas the quantities we calculated in the last section:

$$SS_T = [Y] - [T] = 31{,}136 - 26{,}136.00 = 5000.00$$
$$SS_{S/A} = [Y] - [A] = 31{,}136 - 27{,}312.00 = 3824.00$$
$$SS_A = [A] - [T] = 27{,}312.00 - 26{,}136.00 = 1176.00$$

As a computational check and as a demonstration of the relationship among these three sums of squares, we will apply Eq. (6-1) to these calculations:

$$SS_{S/A} + SS_A = 3824.00 + 1176.00 = 5000.00 = SS_T$$

Comment. There is an alternative way of computing the within-groups sum of squares that illustrates some of the logic underlying the analysis of variance. The $SS_{S/A}$ is actually a *composite* based on the individual sum of squares for each of the treatment groups. That is,

$$SS_{S/A} = SS_{S/A_1} + SS_{S/A_2} + \cdots \qquad (6\text{-}8)$$

In the present case, there are two within-group sums of squares, one for a_1 and one for a_2. Using the data from Table 6-1, we find

$$SS_{S/A_1} = (53^2 + 49^2 + \cdots + 32^2 + 27^2) - \frac{480^2}{12}$$
$$= 20{,}482 - 19{,}200.00 = 1282.00$$
$$SS_{S/A_2} = (47^2 + 42^2 + \cdots + 11^2 + 6^2) - \frac{312^2}{12}$$
$$= 10{,}654 - 8112.00 = 2542.00$$

Completing the operations specified in Eq. (6-8), we find that

$$SS_{S/A_1} + SS_{S/A_2} = 1282.00 + 2542.00 = 3824.00 = SS_{S/A}$$

6.3 MEAN SQUARES AND THE *F* RATIO

We are now ready for the final steps in the analysis: calculating variances and forming the statistic to test for the presence of treatment effects, the *F* ratio. When variances are employed in ANOVA, they are called **mean squares;** the *F* statistic is simply a ratio of two mean squares. A mean square is essentially an "average" sum of squares—not a strict arithmetic average, however, but one based on degrees of freedom rather than on the number of observations.

Degrees of Freedom

As indicated in Chap. 4, the general rule for computing the degrees of freedom (*df*) associated with any sum of squares is

$$df = \begin{pmatrix}\text{number of}\\ \text{independent}\\ \text{observations}\end{pmatrix} - \begin{pmatrix}\text{number of}\\ \text{restraints}\end{pmatrix} \qquad (6\text{-}9)$$

Consider the *df* associated with SS_T. The number of independent observations is $(a)(s)$. There is one restraint placed on this sum of squares, namely, that the sum of the deviations is zero. Thus, $df_T = (a)(s) - 1$.

Consider next the SS_A. In this case, there are *a* independent observations, one for each of the *a* treatment means. The same restraint placed on SS_T is also placed on this sum of squares: the sum of the deviations of the treatment means from the grand mean must equal zero. As a consequence, $df_A = a - 1$, one less than the number of treatment means.

The determination of the *df* associated with the within-groups sum of squares $SS_{S/A}$ is a bit more complicated, but follows the general rule specified by Eq. (6-9). You will recall from Eq. (6-8) that this sum of squares represents a pooling of separate *SS*'s obtained from the different treatment groups. The *df* are obtained the same way. That is, the number of *independent observations* associated with any treatment group is *s* and the *df* for the corresponding *SS* is $s - 1$; the restraint is that the sum of the within-group deviations must sum to zero for *each of the groups*. The $df_{S/A}$ is calculated by combining the separate *df*'s for the different groups:

$$df_{S/A} = df_{S/A_1} + df_{S/A_2} + \cdots$$

Since the *df* for each group is $s - 1$ and there are *a* different groups, we can express the formula as

$$df_{S/A} = (a)(s - 1)$$

Mean Squares

The variance estimates required in ANOVA are given by the formula

$$MS = \frac{SS}{df} \qquad (6\text{-}10)$$

As applied to the two component sources of variance,

$$MS_A = \frac{SS_A}{df_A} \quad \text{and} \quad MS_{S/A} = \frac{SS_{S/A}}{df_{S/A}}$$

The mean square on the left is influenced by two factors, the presence of treatment effects and uncontrolled variation, while the mean square on the right is influenced by uncontrolled variation alone. That is, the MS_A represents the deviation of the treatment groups from the grand mean, which in essence is due to the effects of the independent variable as well as the chance differences that occur in any experiment; the $MS_{S/A}$ represents the deviation of scores within the treatment groups, and since all members of each group are treated alike, it reflects uncontrolled or random variability.

The *F* Ratio

The final step in the calculations is the formation of the *F* ratio, which is used to test for significance (see Chap. 8). For the present type of design, the ratio consists simply of the treatment mean square MS_A divided by the within-groups mean square $MS_{S/A}$:

$$F = \frac{MS_A}{MS_{S/A}} \qquad (6\text{-}11)$$

The result of this division, the *F* statistic, will be used to evaluate the effectiveness of the treatment conditions against the background of uncontrolled, chance factors that are always present and contribute to group differences.

Summary of the Analysis

The computational formulas for the completely randomized single-factor ANOVA are presented in Table 6-2. The first column lists the sources of variance usually extracted from the analysis. Column 2 gives the three basic ratios that are combined in different patterns to produce the sums of squares. These patterns are indicated in column 3. The formulas for the degrees of freedom, mean squares, and *F* ratio are entered in the remaining columns of Table 6-2.

Table 6-2
Summary of the Analysis of Variance

Source	Basic Ratio	*SS*	*df*	*MS*	*F*
A	$[A] = \frac{\Sigma A^2}{s}$	$[A] - [T]$	$a - 1$	$\frac{SS_A}{df_A}$	$\frac{MS_A}{MS_{S/A}}$
S/A	$[Y] = \Sigma Y^2$	$[Y] - [A]$	$(a)(s - 1)$	$\frac{SS_{S/A}}{df_{S/A}}$	
Total	$[T] = \frac{T^2}{(a)(s)}$	$[Y] - [T]$	$(a)(s) - 1$		

The results of the numerical example are summarized in Table 6-3. The *SS*'s were calculated previously and are entered without comment in the table. The *df* for the three sources are found by substituting in the formulas provided in Table 6-2. To be more explicit,

$$df_A = a - 1 = 2 - 1 = 1$$
$$df_{S/A} = (a)(s - 1) = (2)(12 - 1) = (2)(11) = 22$$
$$df_T = (a)(s) - 1 = (2)(12) - 1 = 24 - 1 = 23$$

As a computational check, we can verify that the *df* for the two component sums of squares equals the *df* for the SS_T. That is,

$$df_A + df_{S/A} = 1 + 22 = 23 = df_T$$

The two mean squares are calculated next by dividing each *SS* by the appropriate *df*. For this example,

$$MS_A = \frac{1176.00}{1} = 1176.00 \quad \text{and} \quad MS_A = \frac{3824.00}{22} = 173.82$$

These numbers are entered in the appropriate column of Table 6-3. The *F* ratio is found to be

$$F = \frac{MS_A}{MS_{S/A}} = \frac{1176.00}{173.82} = 6.77$$

The meaning and usefulness of this statistic will be considered in Chap. 8. For now, we will simply note that the ratio of the numerator to the denominator is a rather large value that is not likely to occur when only chance factors are present

Table 6-3
Summary of the Analysis

Source	SS	df	MS	F
A	1176.00	1	1176.00	6.77
S/A	3824.00	22	173.82	
Total	5000.00	23		

6.4 SUMMARY

We have shown how a simple two-group experiment may be analyzed by means of ANOVA. With ANOVA the focus is on differences and deviations. Two sources of variation are critical for the analysis of this design: variation that is attributed to the difference between the two treatments and variation that is not and is uncontrollable. These sources are represented by the sums of squared deviations from means. The SS_A reflects the deviation of the two treatment means from the grand mean. For the within-groups source, the $SS_{S/A}$ reflects the variability of subjects given the same treatment condition. The next step consists of calculating mean squares, which is accomplished by dividing the sums of squares by their appropriate numbers of degrees of freedom. Finally, we calculate the F statistic by dividing the treatment mean square MS_A by the within-groups mean square $MS_{S/A}$. This statistic is used to decide whether the results of the experiment reflect at all the effects of the independent variable. The details of this last step will be discussed in Chap. 8.

6.5 EXERCISES

1. In Sec. 6.1, we showed how the deviation of a particular score (Y_{24}) from the grand mean ($\bar{Y}_T$) may be divided into two components, the deviation of the score from its group mean ($Y_{24} - \bar{Y}_{A_2}$) and the deviation of the group mean from the grand mean ($\bar{Y}_{A_2} - \bar{Y}_T$). These deviations form the basis for the analysis of the results of a single-factor experiment.

a. Calculate the same deviations for *all* the scores in Table 6-1, verifying in each case that the sum of the two component deviations equals the deviation from $\bar{Y}_T$.

b. Square all the deviations for all the subjects and then sum each set of squared deviations over all the subjects. Verify that the three sums are identical to those obtained with the computational formulas in Sec. 6.2 (see Table 6-3).

2. A psychologist decides to determine the effectiveness of a new drug on the ability of rats to learn a difficult maze. The design consists of two conditions, one in which the drug is administered by injection 2 hr before testing and another in which an inert substance (e.g., a saline solution) is substituted for the drug. Each group is represented by $s = 9$ subjects randomly assigned to the conditions. The response measure is the number of trials required to learn the maze according to a criterion of errorless performance. The following data are obtained:

Drug	No Drug
30	36
26	35
31	27
30	32
24	29
28	41
25	36
33	28
31	30

a. Calculate the means and standard deviations for the two treatment groups.
b. Calculate the basic ratios.
c. Find the sums of squares for A, S/A, and T.
d. Determine the degrees of freedom and calculate the mean squares.
e. Construct a summary table and calculate F. (Save this information for Problem 1 at the end of Chap. 8.)

Once we have formulated a researchable hypothesis, then collected data to test that hypothesis, and, finally, calculated statistics to describe and summarize the results, we need to determine whether the difference observed between the two treatment means (or the relationship found between the two variables) is due to the independent variable or whether it is due entirely to chance. To answer this question, we turn to a formal statistical procedure called **hypothesis testing**. We will examine this procedure first within the context of ANOVA and then turn our attention to the significance of correlational statistics. As you will see, both statistical tests result in the same conclusion concerning the outcome of a two-group experiment.

8.1 THE STATISTICAL HYPOTHESES

With ANOVA, we use the F statistic to evaluate the reasonableness of a statistical hypothesis known as the **null hypothesis**, usually symbolized as H_0. The null hypothesis is quite distinct from a research hypothesis, which usually asserts that the treatment conditions will actually produce true differences in performance. In contrast, the null hypothesis usually states that the independent variable in the experiment is completely *ineffective* and that the means associated with the two treatment *populations* (symbolized as μ_1 and μ_2) are *equal*—that is,

$$H_0\text{:}\quad \mu_1 = \mu_2$$

If the difference between the two sample treatment means is too large to be reasonably due to chance factors (and what we mean by "reasonably" we will explain below), the null hypothesis is rejected in favor of a second statistical hypothesis, called the **alternative hypothesis** (H_1). This hypothesis states that the two population treatment means are *not* equal:

$$H_1\text{:}\quad \mu_1 \neq \mu_2$$

A rejection of H_0 leads to the acceptance of H_1, which in effect implies support of our original *research* hypothesis. Failing to reject H_0, on the other hand, can be viewed as a failure of the experiment to support the research hypothesis.

The Logic of the F Ratio

In Chap. 6, we indicated that the F ratio (Eq. 6-11) is written as

$$F = \frac{MS_A}{MS_{S/A}}$$

The denominator of the F ratio, $MS_{S/A}$, provides an estimate of **error variance** regardless of the status of the null hypothesis, and for this reason it is often called the **error term**. To elaborate, you will recall that $MS_{S/A}$ is based on the **variability** of subjects who are *treated alike*, which means that differences between the two treatment groups do not enter into its determination. As a result, $MS_{S/A}$ simply reflects uncontrolled variability—otherwise known as **error variance**—regardless of whether H_0 is true or not.

In contrast, the numerator of the F ratio, MS_A, is based on the difference between the two means and is sensitive to the presence of any treatment effects. Suppose for the moment that the null hypothesis is true. Under these circumstances, any difference observed between the two treatment means must be due to chance factors that result from the random assignment of subjects to groups and other unsystematic factors. This means, then, that MS_A reflects only error variance. On the other hand, when the null hypothesis is *false*, MS_A reflects the *joint operation* of two factors: error variance and treatment effects.

What are the implications of these considerations for the F ratio? First, if the null hypothesis is *true*, the F ratio consists of one estimate of chance factors, based on the chance difference between the two groups, divided by another estimate of chance factors, based entirely on within-group differences. That is, both the numerator and the denominator of the F ratio would contain estimates of experimental error, and we would have

$$\frac{\text{Experimental error}}{\text{Experimental error}}$$

If such an experiment were conducted a large number of times and F ratios were determined for each experimental result, we would expect the average of these F ratios to be approximately 1.0.

On the other hand, if the null hypothesis is *false*, the numerator of the F ratio will be systematically larger than the denominator—on account of the additional presence of treatment effects—and the average F will be *greater* than 1.0. More explicitly, the ratio will become

$$\frac{\text{Treatment effects + experimental error}}{\text{Experimental error}}$$

Unfortunately, the fact that *average* values of F will be different when H_0 holds and when H_1 holds does not really help us in deciding between the two hypotheses in a specific experiment. That is, we must realize that a small F ratio does not guarantee that the null hypothesis is true (and the alternative is false), since chance factors may be counteracting any true difference in the population. By the same token, a large F ratio does not necessarily imply that the null hypothesis is false (and the alternative is true), since large differences between the two groups can occur entirely on the basis of chance. What this means is that we simply cannot be certain of avoiding incorrect decisions under these circumstances—that is, when chance factors are operating in an experiment. The best we can do is to adopt a course of action that minimizes incorrect decisions.

The Sampling Distribution of *F*

Let us see *how* we can minimize the occurrence of incorrect decisions. Suppose we have programmed a computer to draw two samples of scores randomly from a large population and then to compute the *F* statistic. As we have described the situation, the null hypothesis is true, since the two "treatment" means are drawn from the same underlying population. The scores are then "returned" to the population and the procedure is repeated a large number of times. From what we said in the last section, we would expect the mean of the *F*'s to be close to 1.0.[1] A frequency distribution of these *F*'s would provide a picture of the values *F* will take when H_0 is true. Such a distribution is called the **sampling distribution** of the *F* statistic.

Instead of using a computer to generate a sampling distribution of *F*, we can draw the distribution from formulas provided by statistical theory. Consider the sampling distribution of *F* that is presented in Fig. 8-1. This is the theoretical sampling distribution of *F* appropriate for the numerical example we introduced in Chap. 6, namely, two lecture treatment groups with 12 subjects in each group.

As you can see from the figure, extreme values of *F* occur fairly infrequently. Since it is relatively unlikely that large values of *F* will occur when H_0 is true, we adopt the strategy of rejecting the null hypothesis whenever this happens in an actual experiment. All we need is to decide on a definition of "extreme" values of *F*. Once this is done, we can establish a rule to reject H_0 when the *F* from an experiment falls within the region of extreme values, and not to reject H_0 the rest of the time.

Most researchers in the behavioral sciences have adopted a region known as the **5 percent level of significance.** This is an interval that begins with some value of *F*, extends to infinity, and contains the *upper 5 percent* of the distribution. We refer to this value as the **critical value of** *F*. In the present example, the critical value of *F* that divides these extreme values from the rest is $F = 4.30$. That is, 95 percent of the *F*'s have values less than 4.30, while 5 percent have values equal to or greater than 4.30. If we follow the decision rule described in the last paragraph, we would reject H_0 if $F \geq 4.30$ (read "*F* is greater than or equal to 4.30") and not reject H_0 when $F < 4.30$ (read "*F* is less than 4.30"). Stated more formally, the decision rule becomes

Reject H_0 when $F_{observed} \geq 4.30$; otherwise, do not reject H_0.

The *F* distribution is in reality a family of curves; the one appropriate for any experiment is determined jointly by the *df* associated with the numerator and with the denominator terms of the *F* ratio. Since the theoretical sampling distributions are continuous functions, we express the significance level in terms of probability, which is based on the proportion of the total area under the curve associated with the rejection region. The Greek letter α (alpha) is used to symbolize this probability. The 5 percent level of significance is specified by the notation

$$\alpha = .05$$

The *F* Table

The information necessary to determine the critical value of *F*, namely, the beginning of the rejection region, is found in Table A-1 of the Appendix. To use the *F* table, we will need to know three factors: the *df* for the numerator of the *F* ratio, the *df* for the denominator of the *F* ratio, and the significance level we have adopted.

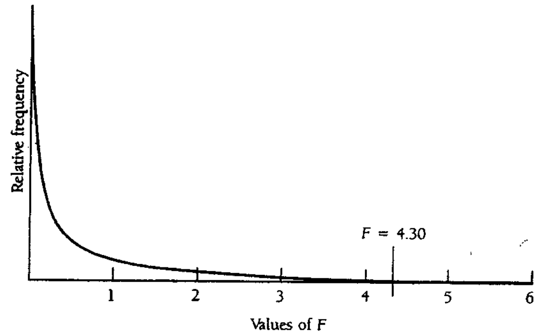


Figure 8-1 Sampling distribution of *F* when there are $a = 2$ groups of $s = 12$ subjects each.

[1] Technically, the *median* of the *F* distribution equals 1. Actually, the mean is slightly larger than 1, since it is defined as

$$\frac{df_{denom.}}{df_{denom.} - 2}$$

This fact does not materially affect the thrust of the argument, however.

In our numerical example presented in Chaps. 6 and 7,

$$df_{num.} = df_A = 1 \qquad \text{and} \qquad df_{denom.} = df_{S/A} = 22$$

We coordinate these two numbers (the column labeled 1 and the row labeled 22) and find critical values of F in Table A-1 for six different significance levels ($\alpha = .25$, .10, .05, .025, .01, and .001). Since we are interested in establishing the rejection region for the 5 percent level of significance, we will use the value listed for $\alpha = .05$, namely, $F = 4.30$. Our decision rule, which we introduced earlier in this section, becomes

Reject H_0 when $F_{observed} \geq 4.30$; otherwise, do not reject H_0.

From Table 6-3, we see that our calculations produced an F of 6.77, which exceeds the critical value of F specified by the decision rule. Consequently, we reject H_0, accept H_1, and conclude that there is a real difference between the two groups such that children receiving the physical science lecture learn more vocabulary words than do children receiving the social science lecture.

We could state this conclusion in other ways. We could say, for example, that the difference between the two groups is significant, or that significantly more vocabulary words are learned when the words are introduced in the context of a physical science lecture than in a social science lecture. The term *significant* is not synonymous with *important*, however. It simply is a shorthand way of stating that the difference observed between the two groups is sufficiently large not to be reasonably attributed to chance factors.

You may have noticed that the *df* values listed in Table A-1 are incomplete. The intervals between successive columns and rows increase with the larger numerator and denominator *df*'s. Fine gradations are not needed for the larger *df* values, however, since the numerical values of F do not change greatly from interval to interval. When the critical value of F falls between two rows or two columns of the table, most researchers follow the practice of choosing the row or column with the *smaller* number of *df*.

Other Significance Levels. Although $\alpha = .05$ is commonly used as the significance level by most researchers, occasionally you will see other probabilities reported in the research literature. As we have already noted, Table A-1 provides F values for six significance levels. In most cases when researchers indicate a probability other than $\alpha = .05$, they are simply providing additional information for readers who may wish to use a different significance level. If you return to the F table (Table A-1), you will see that the calculated F of 6.77 exceeds the critical value of F at $\alpha = .025$ ($F = 5.79$). We could have reported our results as significant at $p < .025$, where p stands for *probability*. This means that the null hypothesis would be rejected by anyone adopting a significance level as small as $\alpha = .025$.

It is important to note that reporting a significance level other than .05 permits *no* inference concerning the strength or magnitude of the effects. We point this out because some researchers have assumed that results that are significant at $p < .025$, for example, are better or stronger than results that are significant at $p < .05$. Comparisons of strength are more appropriately made by obtaining some measure of ***effect magnitude***, which we will discuss in some detail in Chap. 10.

Analyses conducted with computers usually state the **exact probability** of the obtained F statistic. This probability refers to the proportion of the sampling distribution of the F statistic falling at or *above* the F obtained in an experiment. In the present case, for example, $F = 6.77$ has an exact probability of $p = .0163$. Knowing this, a reader can simply apply his or her chosen significance level—for example, $\alpha = .05$—and reject H_0 if the exact probability is smaller (which it is in this example) or not reject H_0 if it is larger. In fact, the decision rule can be stated quite simply, without specific reference to F; that is,

If $p \leq .05$, reject H_0; otherwise, do not reject H_0.

In whatever manner the statistical test is reported, however, we must not forget that *our* significance level is decided upon *before* the start of an experiment and alternative ways of reporting probabilities do not change this fundamental point.

8.2 TESTING THE SIGNIFICANCE OF A CORRELATION

As you have seen, the final step in an analysis of variance is to evaluate the significance of the F statistic. What is evaluated is the null hypothesis, which states that the population treatment means are equal. If the $F_{observed}$ falls within the rejection region, we conclude that treatment effects are present—that is, that the population treatment means are not the same. We evaluate the statistical significance of the r statistic in a similar fashion. In this case, the null hypothesis states that there is no relationship between X and Y in the population. In symbols,

$$H_0\text{:} \quad \rho = 0$$

where ρ (Greek letter rho) represents the correlation between X and Y in the population. If this hypothesis is rejected, by applying procedures we will describe in the next paragraph, we accept the alternative hypothesis,

$$H_1\text{:} \quad \rho \neq 0$$

and conclude that there is a nonzero relationship between the two variables. In the context of the present example, we would conclude that there is a significant association between the type of lecture (the X variable) and vocabulary test scores (the

Y variable). As we have stated before, this conclusion—that there is a *relationship*—is identical to the conclusion that there is a *difference* between the treatments in the number of words learned.

As already pointed out in Chap. 7, an appropriate *F* ratio is given by Eq. (7-1):

$$F = \frac{r^2/df_{reg.}}{(1 - r^2)/df_{res.}}$$

The *df* in the numerator ($df_{reg.}$) is equal to 1, reflecting the fact that the *df* for a two-treatment experiment is 1. The *df* in the denominator ($df_{res.}$) is equal to $N - 2$, for reasons we will discuss in a moment. For convenience, we can rewrite Eq. (7-1) as follows:

$$F = \frac{r^2}{(1 - r^2)/(N - 2)} \tag{8-1}$$

Entering the appropriate data into this equation, we find

$$F = \frac{(.4850)^2}{[1 - (.4850)^2]/(24 - 2)}$$
$$= \frac{.2352}{(.7648)/22} = \frac{.2352}{.0348} = 6.76.$$

This *F* is evaluated in the usual manner with

$$df_{num.} = 1 \quad \text{and} \quad df_{denom.} = N - 2 = 24 - 2 = 22$$

Since the $F_{observed}$ is greater than 4.30 (the critical value for $\alpha = .05$, when $df = 1, 22$), we can declare that the correlation between the type of lecture and vocabulary scores is significant. You should note that the value of $F_{observed}$ is identical within rounding to that calculated with ANOVA (6.77).

Comparison with ANOVA

In order to gain further insight into the parallels between ANOVA and MRC, we can explore and dissect the two *F* ratios used in the statistical evaluation of the corresponding null hypotheses. Suppose we express Eq. (8-1) in terms of sums of squares. This is easily accomplished by recalling from Eq. (5-15) that

$$r^2 = \frac{SS_{reg.}}{SS_Y}$$

and from our discussion in Chap. 5 that

$$1 - r^2 = \frac{SS_{res.}}{SS_Y}$$

and then substituting this information in Eq. (8-1). The final result of substitution is

$$F = \frac{SS_{reg.}}{SS_{res.}/(N - 2)} \tag{8-2}$$

What do the components in Eq. (8-2) represent? Let us begin with the denominator, $SS_{res.}/(N - 2)$. Of particular relevance is the term $N - 2$, which reflects the number of independent observations (the total number of subjects) and the value 2. This value 2 represents the two restraints or restrictions placed on the calculations. These result from the process of obtaining the predicted *Y* values (*Y'*) from the regression equation, which uses two pieces of information in order to estimate the slope *b* and the intercept *a* of the regression line from the *N* pairs of observations available. From the ANOVA framework, the corresponding sum of squares is $SS_{S/A}$, where the restrictions result from estimating the two group means $\bar{Y}_{A_1}$ and $\bar{Y}_{A_2}$ upon which the deviations of the two sets of *Y* scores are based. The degrees of freedom in this latter case are indicated as $df_{S/A} = (a)(s - 1)$, where *a* is the number of treatment groups and *s* is the number of subjects assigned to each treatment group.

Both of these sources of variance have the same number of degrees of freedom. That is,

$$df_{res.} = N - 2 = 24 - 2 = 22$$
$$df_{S/A} = (a)(s - 1) = (2)(12 - 1) = 22$$

In other words, the quantity $N - 2$ represents the degrees of freedom associated with $SS_{res.}$, and thus, the denominator of Eq. (8-2) is a mean square—in this case, a mean square for the residual deviations found with linear regression ($MS_{res.}$). That is,

$$\frac{SS_{res.}}{N - 2} = \frac{SS_{res.}}{df_{res.}} = MS_{res.}$$

Though we indicated that the equation for the *F* ratio will be different for experiments in which there are more than two treatments, the conceptual underpinnings are the same. That is, the error term in the denominator will include the *df* for the error term in the experiment.

Now let us look at the numerator of Eq. (8-2). The *df* associated with $SS_{reg.}$ is 1; this sole *df* represents the deviation of the slope constant *b* from zero. Thus, the numerator could have been written as

$$\frac{SS_{reg.}}{1} = \frac{SS_{reg.}}{df_{reg.}} = MS_{reg.}$$

In sum, the F value obtained to test the significance of a correlation coefficient is a ratio of the $MS_{reg.}/MS_{res.}$, with $df_{num.} = 1$ and $df_{denom.} = N - 2$. This is equivalent to the F ratio of ANOVA for a two-treatment condition, which is $F = MS_A/MS_{S/A}$. In correlational analysis, the denominator reflects an error term against which the effect, $MS_{reg.}$, in the numerator is tested; similarly, in ANOVA, the denominator reflects experimental error and the numerator reflects a combination of treatment effect and experimental error.

8.3 TESTING THE SIGNIFICANCE OF *b* IN THE REGRESSION EQUATION

How does the b value relate to r in terms of testing a specific hypothesis in a two-group design? Since we have shown in Sec. 7.2 that the regression coefficient based on contrast coding represents information about the difference between the two means, it makes sense to ask whether the value is significantly different from zero. An F test for assessing the significance of the regression coefficient is given by[2]

$$F = \frac{b^2}{MS_{res.}/\Sigma\,(X - \bar{X})^2} \tag{8-3}$$

For the data in Table 7-1, we previously calculated the following quantities:

$$b = 7.00 \qquad S_{res.} = 3824.00 \qquad \Sigma\,(X - \bar{X})^2 = 24.00$$

Thus,

$$MS_{res.} = \frac{SS_{res.}}{df_{res.}} = \frac{3824.00}{22} = 173.82$$

Applying Eq. (8-3) yields:

$$F = \frac{7.00^2}{173.82/24.00} = \frac{49.00}{7.24} = 6.77$$

which is equal to the F we obtained when we conducted the test of significance of r. In other words, a test of b, which *reflects* the relationship between X and Y, is identical to a test of r, which is the index of the relationship between X and Y. Furthermore, these tests equal the ANOVA test of the difference between the two means for the two levels of X.

[2] This test is usually expressed as a t test (see Edwards, 1976, p. 106; and McNemar, 1969, p. 160). In this form, the denominator is the standard error of the regression coefficient.

8.4 SOME THEORETICAL CONSIDERATIONS

In this section, we will first consider some problems that are inherently associated with testing hypotheses regardless of whether the context is ANOVA or MRC. Next, we will discuss a related topic, the sensitivity of a statistical test in detecting differences in the population. We will conclude with a brief consideration of the statistical models underlying ANOVA and MRC.

Errors of Statistical Inference

It must be realized at the outset that conclusions drawn from *any* statistical test may be in error. We must make this dismal pronouncement because we have no way of determining the exact situation existing in the treatment populations. All that we have are the data from an experiment, which are assumed to consist of random samples drawn from the different treatment populations. Consequently, any conclusion we may extract from our data represents what is in effect an educated guess about these unknown treatment populations.

What this means, therefore, is that we have no sure way of avoiding errors of statistical inference and, moreover, that we will never know when we are committing them! Realistically, all that we can do is to take steps to *minimize* the occurrence of such errors. We represent our calculated risk taking in terms of probabilities, but in speaking of probabilities we emphasize the susceptibility to error of all conclusions based on sample data. We will consider two types of error: an error we may commit whenever we reject the null hypothesis, and another we may commit whenever we do not reject the null hypothesis. These are known as **type I error** and **type II error**, respectively.

Type I Error. As we have implied, we make a type I error whenever we falsely reject the null hypothesis—whenever we conclude that differences exist among the population treatment means or that the correlation is greater than zero when in fact such is not the case. We directly control the probability of this error in our choice of *significance level*. The probability specified by $\alpha = .05$, for example, refers to the proportion of F's that theoretically occur beyond a particular point on the F distribution—in this case, the value of F marking off the 95th percentile ($F_{.95}$). Under the null hypothesis, values of F falling within this area will occur by chance 5 percent of the time. If we accept $\alpha = .05$ and obtain an F that equals or exceeds the value of $F_{.95}$, we assume that this is *not* one of the extreme F's expected to occur by chance and conclude instead that the null hypothesis is false. But note that we are making an assumption—that the F has not occurred by chance—and that is the very reason why type I errors are committed. This also explains why we set α at .05 or smaller: to keep the probability of our committing this type of error at a reasonably low level. If we set $\alpha = .05$, then, we agree in principle to make a type I error 5 percent of the time when the null hypothesis is true.

*

Type II Error. In the preceding section we focused on the possibility of rejecting the null hypothesis, when in reality the treatment means are *equal* or the correlation is *zero*. What about the more interesting situation when the means are *not* the same or the correlation is different from zero but we make the error of failing to reject the null hypothesis? When this happens, we commit a type II error. The probability associated with this error is represented by the Greek letter β (beta). Unfortunately, we will never know the exact probability of this error, even though we have a Greek letter reserved for it! This is because we must possess certain details about the theoretical treatment populations—the means and standard deviations or correlation—in order to determine this probability. The best we can do is to make certain assumptions about these parameters and take steps that are known to keep β at a reasonably low value. We will discuss these steps in a moment.

Comment. We always face the possibility of making an error whenever we draw conclusions from a set of data. The type of error depends on the nature of the conclusion drawn from the results. If the population treatment means really are the same, we will make an error if we conclude that a difference exists between them; this is known as a type I error. The probability of a type I error is specified by our choice of significance level. On the other hand, if in fact a difference exists between the two treatment populations, we will make an error if we fail to reject the null hypothesis; this is known as a type II error.

You should note that we can make only one of these errors in any given statistical test—not both. This is because each error is exclusively associated with a *different conclusion*: a type I error is associated with *rejecting* the null hypothesis, and a type II error with *not rejecting* the null hypothesis. Since the decision rule forces us to choose one of these two conclusions, we will be susceptible only to the error associated with the particular conclusion we make.

Power of a Statistical Test

Power is a statistical concept that refers to the probability of *correctly* rejecting the null hypothesis. It is related to the probability of a type II error as follows:

$$\text{Power} = 1 - \beta \qquad (8\text{-}4)$$

If the probability of making a type II error is .30, for example, power is 1 − .30 = .70. What this means is that if we repeated the same experiment over and over, we would reject the null hypothesis 70 percent of the time; 30 percent of the time we would commit a type II error.

In contrast with type I error, which is controlled directly by our choice of α, type II error (or power) is controlled indirectly. Several factors are known to increase power, but the most common way is to increase sample size s. There is a practical limit to the use of this strategy, however, since experiments with already large sample sizes require quite sizable increases in sample size to achieve the same gain in power that one would obtain by adding subjects to a less ambitious experiment. An experiment with a sample size of $s = 5$, for example, will benefit more by the addition of another 5 observations per group than will an experiment with $s = 10$ or $s = 20$.

Another way of increasing power consists of bringing under closer control any unsystematic sources of variability that may be operating in an experiment. Most commonly, this involves reducing subject variability through the use of a more homogeneous pool of subjects. Alternatively, more sensitive experimental designs might be chosen, such as designs which use either subjects who are matched on some relevant ability or factor or subjects who serve in *all* the treatment conditions instead of only one. Both of these alternative designs can result in a sizable increase in power, but require a type of statistical analysis other than we have considered so far. A final method, called the **analysis of covariance**, provides a *statistical* solution whereby information about the subjects is collected before the start of the experiment and then used to reduce the influence of chance factors in the experiment. We will consider these alternative designs and the analysis of covariance in later chapters.

Most experiments in the behavioral sciences are designed without consideration of power and, often, are seriously lacking in their ability to detect differences among treatment means when they are present in the population. Cohen (1962) and Brewer (1972) have observed astonishingly low levels of power for experiments reported in the psychological literature. What this means is that findings associated with lower power stand a poor chance of being duplicated by others who wish to repeat or to extend these studies. For this reason, then, we feel it is reckless not to obtain power estimates in the planning stage of an experiment. It is at this point that something can still be done to increase power if it is too low, either by increasing sample size, by attempting to reduce error variance, or by choosing a potentially more sensitive experimental design or statistical procedure. Power determinations are not difficult to obtain, and they often provide useful insight into the nature of the phenomenon under study. Cohen devotes an entire book (1977) to the discussion of power in a variety of different research settings. Other, less comprehensive presentations are also available; see for example, Keppel (1982, Chap. 4), Myers (1979, pp. 86–88), and Winer (1971, pp. 220–228). We suggest that you consult these references for further discussion of the topic

Statistical Models and Assumptions

The theoretical justification of the inferential procedures we have outlined in this chapter is dependent on a set of assumptions and complex statistical proofs. We will highlight some of these arguments here.

The Analysis of Variance. Underlying ANOVA is a model—known as the **linear model**—that expresses the score of a subject in any treatment condition as the sum of certain parameters of the population. Three assumptions underlie the use of the F distribution for evaluating the null hypothesis. Briefly, we assume that the treatment populations are *normally distributed*, that they have *equal variances*, and that the individual observations are *independent* of (i.e., uninfluenced by) any other observations, either within the same treatment population or between treatment populations.

Research over the last two decades has shown that even sizable violations of the first two assumptions do not appear to distort the distribution of the F statistic seriously.[3] The assumption of independence is usually satisfied by assigning subjects randomly to the treatment conditions and administering the treatments individually to the different subjects. In short, the F statistic is amazingly insensitive to even flagrant violations of the assumptions of normality and of the homogeneity of variances. The main requirement is that subjects be randomly assigned to the treatments.

Multiple Regression and Correlation. The model underlying the MRC analysis is algebraically equivalent to the one underlying ANOVA. The difference is in the way the linear model is expressed. The assumptions underlying the use of the F distribution in evaluating the null hypothesis are identical to those summarized for ANOVA. That is, it is assumed that the treatment populations are normally distributed and have equal variances and that all individual observations are independent of one another.

8.5 SUMMARY

The final step in the statistical analysis of an experiment is the significance test. Significance testing begins with the formulation of a *null hypothesis* (H_0), which will be evaluated with the statistical evidence generated by the study. This hypothesis generally consists of a statement proposing that there are *no treatment effects* or that there is *no relationship between the variables*. With ANOVA, for example, the null hypothesis states that the two population treatment means are the same; that is,

$$H_0: \quad \mu_1 = \mu_2$$

With correlational analysis, the corresponding null hypothesis states that the correlation between the independent variable X and the dependent variable Y in the population is zero. In symbols,

$$H_0: \rho = 0$$

(An alternative approach is to test the significance of the regression coefficient; in this case, the null hypothesis would state that the slope of the regression line relating X and Y in the population is zero.) A second statistical hypothesis, which is called the *alternative hypothesis* (H_1), is also formulated at this time. This hypothesis essentially states that the null hypothesis is *false*, implying that there is a difference between the population treatment means or a correlation in the population between X and Y. Our task now is to decide which of these two statistical hypotheses is more likely to be correct, given the outcome of the experiment we have just completed.

At this point, we return to the experiment and, depending on the statistical approach we have followed, examine either the observed difference between the two means (for ANOVA) or the correlation between X and Y (for MRC). Because of the operation of chance factors, which stem largely from the random assignment of subjects to conditions, we fully expect to find some difference between the two means or a nonzero value for the product-moment correlation even if the null hypothesis is true. To deal with this problem, we calculate an F ratio which relates *systematic variation* (variation associated with the experimental manipulation) to *unsystematic variation* (chance or random variation). Systematic variation, which is reflected by MS_A in ANOVA and by r^2 in MRC, is influenced by two sources, chance effects and potential treatment effects; while unsystematic variation, which is reflected by $MS_{S/A}$ in ANOVA and by $(1 - r^2)/(N - 2)$ in MRC, is influenced by chance factors alone.

We now compare the value of F obtained in the experiment with the so-called *critical value* of F, which is based on the theoretical sampling distribution of F and is found in a statistical table. This value sets the lower boundary of the range of F's within which we will reject the null hypothesis. If the observed F falls within this range—if it is equal to or greater than the critical value of F—we reject the null hypothesis and conclude that treatment effects are present in the population. If the observed F is smaller than the critical value, we do not reject the null hypothesis.

[3] The F statistic is sensitive to concurrent or simultaneous violations of the assumptions of normality and homogeneity. See Myers (1979, pp. 66–72) for an excellent discussion of these problems.

We discussed in detail the errors of statistical inference that may occur through hypothesis testing. A type I error occurs when a null hypothesis is rejected falsely; we keep the probability of such errors at a low value through our choice of significance level. A type II error occurs when treatment effects are present in the population but the null hypothesis is not rejected; we control the probability of such errors indirectly through our choice of sample size and of experimental design. Power, which is defined in terms of type II error, refers to the sensitivity of a statistical test. A consideration of the statistical models underlying ANOVA and MRC reveals that the F test is relatively insensitive to violations of the assumptions of normality and of homogeneous treatment variances.

8.6 EXERCISES

1. Consider again the experiment presented in Problem 2 of Chap. 6, which you analyzed with analysis of variance.

a. State the decision rule for rejecting the null hypothesis, using $\alpha = .05$. Is the F obtained from ANOVA significant?
b. What is the decision rule if instead you use $\alpha = .10$? Is the F significant?

2. In Problem 1, Chap. 7, you calculated an r based on this same set of data.

a. Complete the correlational analysis by calculating the F ratio.
b. State the decision rule for evaluating the null hypothesis at $\alpha = .05$. Is the F significant?
c. Verify that the F from ANOVA and that from the correlational analysis are the same.

9

General Coding of Experiments for MRC Analysis

. . .

9.1 THE CODING OF TREATMENT CONDITIONS
A General Rule for Coding
Three Types of Coding
9.2 THE CODE MATRIX
9.3 SUMMARY
9.4 EXERCISES

You have seen how the analysis of a single-factor experiment by either ANOVA or MRC is easily extended to any number of treatment conditions. Up to this point, we have considered only one statistical test—the evaluation of the omnibus null hypothesis (indicated by the F ratio for MS_A or for $R^2_{Y.max.}$)—which in most cases is *not* of primary interest to the researcher. This is because the rejection of this null hypothesis simply tells us that it is reasonable to conclude that there is a relationship (MRC) or that differences between treatment conditions are present (ANOVA), but the rejection does not tell us *which* treatments are different. As a consequence, researchers usually plan an experiment around a limited number of focused comparisons that *will* indicate exactly which aspects of the independent variable are producing significant differences and which are not.

The nature of the analyses we might consider for any experiment generally depends on the *research hypotheses* that guided us in the selection of the treatment conditions to be included in the study. Quite naturally, the statistical analysis consists of comparisons created by grouping different subsets of treatment means that in turn provide answers to these questions. If our independent variable consists of *qualitative* manipulations, the analysis will usually take the form of comparisons between pairs of means. On the other hand, if it consists of a *quantitative* manipulation, the analysis will probably focus on an attempt to identify the underlying *trend* or *shape* of the relationship between the independent and dependent variables. This chapter deals with the analysis of qualitative independent variables, which are frequently called **categorical** or **nominal independent variables**. In Chap. 23 we examine the analysis of trend.

11.1 PLANNED COMPARISONS

Planned comparisons are analyses that are planned *before* the start of the experiment. They are frequently obtained by translating research hypotheses into comparisons between means from selected treatments. Usually, these comparisons are tested directly *without* any preliminary assessment of the omnibus F test. As you will see, planned comparisons offer an analytically powerful approach to the analysis of an experiment.

Types of Planned Comparisons

In an experiment with more than two treatment conditions, the most common planned comparison consists simply of the difference between two means. Typically, this difference will be based on a comparison of the mean of one treatment group with the mean of another treatment group. Such a comparison is often called a *pairwise comparison* because it is based on a difference between a pair of treatment means. Less commonly, researchers use *complex* comparisons in which one or even both of the two means being compared are themselves averages of two or more treatment means. We will consider examples of both types of comparisons—pairwise and complex—in a moment.

Either type of comparison is often called a **single-*df* comparison**, in reference to the single degree of freedom associated with a difference between two means. Viewed another way, a single-*df* comparison is equivalent to the treatment source of variation (factor A) obtained from an experiment containing $a = 2$ treatment groups, in which the *df* for the treatment source is 1.

As an illustration of planned comparisons, let us return to our continuing example in which vocabulary words were introduced through lectures dealing with physical science, social science, or history. Consider the information provided by this particular experiment. There are three differences between pairs of means—pairwise comparisons—that we might examine, namely, physical science versus social science, physical science versus history, and social science versus history. The original two-group experiment, which we considered first in Chap. 6, yielded only the first difference.

In addition, we might examine at least one complex comparison, the difference between an average of the two science lectures and the history lecture. Two other complex comparisons that are possible with this design, an average of physical science and history versus social science and an average of social science and history versus physical science, do not provide as sharp a comparison as the first and probably would not be of interest to a researcher. In general, the quality of a complex comparison depends on the *logical* basis for averaging treatment conditions. In the first case, taking the average of the physical and social sciences and comparing it with the history condition is based on commonalities between the two sciences not shared with history; in the other two cases, the basis for the comparison is less obvious.

One other type of comparison is common in the behavioral sciences. This occurs when there is a subset of logically similar treatment conditions included as part of a larger study. Suppose we included a fourth condition—a lesson on biological science—in our growing vocabulary experiment. In addition to a variety of new meaningful single-*df* comparisons afforded by this expanded design, it is also possible to consider the differences between means within the subset of *science conditions* (physical, social, and biological sciences). The degrees of freedom for this subset are 1 less than the number of means being examined; that is, $df_{A_{set}} = 3 - 1 = 2$. Common examples of this type of analysis are also found in experiments containing a control group and several experimental groups. In these cases, researchers typically assess the differences within the set of experimental groups, as well as a number of interesting single-*df* comparisons.[1]

The Omnibus *F* Test

The *F* test we considered in Secs. 10.1 and 10.2—the omnibus *F* test—was a single statistical test assessing either significance of the differences among *all* the treatment means or the overall association between type of lecture and vocabulary words learned. This test does not tell us, however, which of these differences are significant and which are not. The omnibus *F* test evaluates what in effect is an *average* of all possible pairwise comparisons.

When we plan specific comparisons, we are not generally interested in the outcome of the omnibus test. Indeed, there is no logical need to conduct the test at all! With planned comparisons, our interest is in certain comparisons and *not* in an average of all pairwise differences. On the other hand, without specific comparisons (or research hypotheses) to guide the analysis—and specific comparisons may be lacking in certain exploratory work—we would probably conduct the omnibus test *first* and let the outcome of the test determine whether we examine the data in more detail.

Such a situation might occur, for example, if we were comparing a number of alternative procedures or products with the goal of identifying the best (or the worst) from the entire set. Under these circumstances, then, the omnibus test tells us whether it is reasonable to conclude that the population treatment means are not all the same. If there is insufficient evidence to reject the omnibus null hypothesis, we conclude that the differences among the treatment means are most likely the result of chance factors that are present in any experiment. On the other hand, if we reject the null hypothesis, we conclude that the means are not all the same and follow the omnibus test with a systematic examination of the data in order to locate the specific differences between the treatment means that are responsible for the significant omnibus *F*. Again, we must stress that rejecting the overall null hypothesis does not identify which means are the same and which are different. Additional analyses are necessary to obtain that important information.

Most experiments in the behavioral sciences are designed to test specific hypotheses, however, and, in our opinion, should be evaluated directly, without reference to the omnibus test. We emphasize this point because one frequently encounters experiments in the research literature that report the result of the omnibus test first, followed by what are in effect *planned comparisons*. We suspect that in most cases the inclusion of the omnibus test is a habit the experimenter acquired when this two-step procedure was in common use.

Post Hoc Comparisons

Post hoc comparisons refer to comparisons conducted *after* the data have been assessed by an omnibus *F*. Post hoc comparisons are *unplanned* in the sense that they are suggested by the outcome of the experiment and are not specifically anticipated during the planning stage of the research project. In most cases, they consist of comparisons following up the results of the major analyses. Such comparisons should possess the same qualities associated with planned comparisons: they should be analytical, and they should make sense.

Post hoc comparisons are sometimes called **multiple comparisons**, a somewhat derogatory term that generally refers to the indiscriminate examination of *all possible comparisons*—usually pairwise differences—in an attempt to locate significant effects. There are special procedures available for dealing with multiple comparisons. We believe, however, that most researchers should restrain themselves and focus their attention only on those comparisons that are meaningful and relevant to the original questions guiding the investigation. We will elaborate this point in the next chapter.

11.2 SINGLE-*df* COMPARISONS: THE ANOVA APPROACH

Planned comparisons permit researchers to ask highly focused questions of a set of data. Generally, such questions are expressed as differences between two means, and take the form of single-*df* comparisons. The two means can be means from specific treatment conditions or means formed by combining treatment conditions. We will start by demonstrating how to calculate a "weighted" difference between means and then show how easily the statistical test can be performed.

Single-*df* comparisons are conducted in two steps: first, calculating the difference between the two means of interest; and, second, evaluating the significance of the difference. While you should have no difficulty in calculating the difference, you probably have no clear idea how to translate this difference into a form that then can be used to form an *F* ratio. To facilitate this latter calculation, we will introduce a procedure which at first will appear to obscure the process of calculating the desired difference, but which will simplify the calculation of the quantities needed for the statistical test.

[1] See Keppel (1982, pp. 123–124) for a more detailed discussion of this type of analysis.

Expressing a Difference as a Sum of Weighted Means

It is useful to express a single-*df* comparison as the *sum* of all the means taken after each has been multiplied by a special weight (or **coefficient**, as it is called). Consider the following expression:

$$\hat{\psi} = (c_1)(\bar{Y}_{A_1}) + (c_2)(\bar{Y}_{A_2}) + (c_3)(\bar{Y}_{A_3}) + \cdots \qquad (11\text{-}1)$$

where $\hat{\psi}$ (Greek psi) = the difference obtained from a given comparison

c_1, c_2, c_3 = the coefficients (or weights) assigned to the treatment means in the experiment

$\bar{Y}_{A_1}, \bar{Y}_{A_2}, \bar{Y}_{A_3}$ = the corresponding treatment means

A compact way of expressing Eq. (11-1) is as follows:

$$\hat{\psi} = \Sigma\,(c_i)(\bar{Y}_{A_i}) \qquad (11\text{-}2)$$

The critical ingredient in these two formulas is the coefficients, which we create to represent a particular comparison.

Let us see how this procedure works. Suppose we wanted to compare the means of the two science conditions. We accomplish this without using coefficients simply by subtracting one mean from the other, that is, taking $\bar{Y}_{A_1} - \bar{Y}_{A_2}$ (or $\bar{Y}_{A_2} - \bar{Y}_{A_1}$; the two differences have the same value except for the sign). We may express this difference using Eq. (11-1), by assigning the coefficient $c_1 = +1$ to physical science, $c_2 = -1$ to social science, and $c_3 = 0$ to history. Entering these coefficients in Eq. (11-1), we find

$$\begin{aligned}\hat{\psi} &= (+1)(\bar{Y}_{A_1}) + (-1)(\bar{Y}_{A_2}) + (0)(\bar{Y}_{A_3})\\ &= \bar{Y}_{A_1} - \bar{Y}_{A_2} + 0\\ &= \bar{Y}_{A_1} - \bar{Y}_{A_2}\end{aligned}$$

As you will soon see, the advantage of expressing this difference between the two science conditions as a sum of weighted means is that this difference ($\hat{\psi}$) and the three coefficients (+1, −1, 0) provide all the information we need to translate the difference into a sum of squares and then into a mean square and an *F* ratio.

For comparisons between pairs of treatment means, the coefficients are simple to express:

+1 and −1 for the two means being compared
0 for all other treatment means in the experiment

For more complex comparisons, which involve means based on combinations of treatment conditions, the coefficients must be created individually for each comparison, a process we will consider next.

In general, the coefficients may be derived from a specification of the actual difference under consideration. To illustrate, suppose we performed an experiment with $a = 6$ treatment conditions and wanted to compare the average of the means for groups 1 and 2 with the average of the means for groups 3, 4, and 5. This comparison, expressed as a difference between two means, becomes

$$\hat{\psi} = \frac{\bar{Y}_{A_1} + \bar{Y}_{A_2}}{2} - \frac{\bar{Y}_{A_3} + \bar{Y}_{A_4} + \bar{Y}_{A_5}}{3}$$

Rewriting this expression slightly, we have

$$\begin{aligned}\hat{\psi} &= (+\tfrac{1}{2})(\bar{Y}_{A_1} + \bar{Y}_{A_2}) + (-\tfrac{1}{3})(\bar{Y}_{A_3} + \bar{Y}_{A_4} + \bar{Y}_{A_5})\\ &= (+\tfrac{1}{2})(\bar{Y}_{A_1}) + (+\tfrac{1}{2})(\bar{Y}_{A_2}) + (-\tfrac{1}{3})(\bar{Y}_{A_3}) + (-\tfrac{1}{3})(\bar{Y}_{A_4}) + (-\tfrac{1}{3})(\bar{Y}_{A_5})\end{aligned}$$

From the original expression of the difference between two means, we easily determine that the coefficient is $+\frac{1}{2}$ for the two groups contributing to the first average (groups 1 and 2), $-\frac{1}{3}$ for the three groups contributing to the second average (groups 3, 4, and 5), and 0 for the one group not entering into the comparison (group 6). Thus, the set of coefficients—

$$+\tfrac{1}{2},\quad +\tfrac{1}{2},\quad -\tfrac{1}{3},\quad -\tfrac{1}{3},\quad -\tfrac{1}{3},\quad \text{and}\quad 0$$

—which will be used in subsequent calculations, can be obtained directly from the original mathematical expression representing a specific difference between two sets of means.

To illustrate further, suppose we had the following treatment means: 15, 20, and 30. If we wanted to examine a pairwise comparison between condition 1 and condition 3, the difference would be $\bar{Y}_{A_1} - \bar{Y}_{A_3} = 15 - 30 = -15$. A set of coefficients representing this difference is +1, 0, and −1, which when substituted in Eq. (11-1) produces

$$\begin{aligned}\hat{\psi} &= (+1)(\bar{Y}_{A_1}) + (0)(\bar{Y}_{A_2}) + (-1)(\bar{Y}_{A_3})\\ &= (+1)(15) + (0)(20) + (-1)(30)\\ &= 15 - 30 = -15\end{aligned}$$

or the same value we obtained by subtracting the third mean ($\bar{Y}_{A_3}$) from the first mean ($\bar{Y}_{A_1}$). As another example, suppose we wanted to contrast condition 1 with the average of conditions 2 and 3. In this case,

$$\begin{aligned}\hat{\psi} &= \bar{Y}_{A_1} - \frac{\bar{Y}_{A_2} + \bar{Y}_{A_3}}{2}\\ &= 15 - \frac{20 + 30}{2} = 15 - 25 = -10\end{aligned}$$

A set of coefficients representing this difference is $+1$, $-\frac{1}{2}$, and $-\frac{1}{2}$. These coefficients are easier to see if we represent the difference as

$$\hat{\psi} = \frac{\bar{Y}_{A_1}}{1} - \frac{\bar{Y}_{A_2} + \bar{Y}_{A_3}}{2}$$

where the coefficient for $\bar{Y}_{A_1}$ is $\frac{1}{1} = +1$ and the coefficients for $\bar{Y}_{A_2}$ and $\bar{Y}_{A_3}$ are both $-\frac{1}{2}$. Substituting in Eq. (11-1), we find

$$\begin{aligned}\hat{\psi} &= (+1)(15) + (-\tfrac{1}{2})(20) + (-\tfrac{1}{2})(30)\\ &= 15 - 10 - 15 = -10\end{aligned}$$

For any single-*df* comparison, then, each treatment mean has a coefficient chosen to reflect the difference under consideration. For all means entering into the comparison, each coefficient has a numerical value and a sign; for means not entering into the comparison, the coefficient is 0. *An important property of a set of coefficients is that they sum to zero.* That is,

$$\Sigma\, c_i = 0 \tag{11-3}$$

In the example with the six conditions,

$$\begin{aligned}\Sigma\, c_i &= (+\tfrac{1}{2}) + (+\tfrac{1}{2}) + (-\tfrac{1}{3}) + (-\tfrac{1}{3}) + (-\tfrac{1}{3}) + (0)\\ &= (+1) + (-1) = 0\end{aligned}$$

Coefficients generated by the method we have just outlined represent what may be called a **standard set** of coefficients. As a matter of fact, equivalent sets can be derived from the standard set simply by multiplying all coefficients in the set by a constant; coefficients obtained this way will produce exactly the same numerical outcome for the statistical test. Researchers often take advantage of this property to transform fractional coefficients into more convenient whole numbers; this is done by multiplying the standard set of coefficients by the lowest common denominator of the set—that is, the smallest value that can be divided by *both* values in the denominator of the coefficients. For example, if we multiplied the coefficients for the complex comparison we derived previously by 6, which is the smallest number divisible by both 2 and 3, we would produce the set $(+3, +3, -2, -2, -2$, and $0)$. One advantage of standard sets, however, is that the value of the comparison itself $(\hat{\psi})$ is always expressed as a difference between two means, which is useful when a researcher wishes to construct confidence intervals based on single-*df* comparisons (see Myers, 1979, pp. 305–306). Coefficients obtained by multiplying the standard set by a constant cannot be conveniently used for this purpose, because the value of $\hat{\psi}$ will also reflect the multiplication.

Sums of Squares

Computational Formula. A single-*df* comparison can be easily translated into a sum of squares. We start with the observed difference between the two means $(\hat{\psi})$. By combining this difference with two other familiar quantities, namely, the coefficients c_i and the sample size s, we can now calculate the sum of squares corresponding to any single-*df* comparison as follows:

$$SS_{comp.} = \frac{(s)(\hat{\psi})^2}{\Sigma\,(c_i)^2} \tag{11-4}$$

Numerical Examples. As a simple illustration of Eq. (11-4), consider the numerical example from Chap. 6 in which only the two science lessons were compared. (We will present a more complex illustration in a moment.) From Table 6-1, we find that the mean for the physical science condition is $\bar{Y}_{A_1} = 40.00$ and the mean for the social science condition is $\bar{Y}_{A_2} = 26.00$; there are $s = 12$ children in each group. The coefficients representing the difference between these two means are $+1$ and -1, respectively. From Eq. (11-1), we find

$$\begin{aligned}\hat{\psi} &= (c_1)(\bar{Y}_{A_1}) + (c_2)(\bar{Y}_{A_2})\\ &= (+1)(40.00) + (-1)(26.00) = 40.00 - 26.00 = 14.00\end{aligned}$$

Substituting in Eq. (11-4), we calculate

$$\begin{aligned}SS_{comp.} &= \frac{(s)(\hat{\psi})^2}{\Sigma\,(c_i)^2}\\ &= \frac{(12)(14.00)^2}{(+1)^2 + (-1)^2}\\ &= \frac{(12)(196)}{1+1} = \frac{2352.00}{2} = 1176.00\end{aligned}$$

You will note that this sum of squares is identical to the SS_A we obtained in Chap. 6 (see Table 6-3). In fact, this is why we chose this example, namely, to demonstrate that Eq. (11-4) and the general formula for calculating the between-groups sum of squares are equivalent when both are applied to an experiment with $a = 2$ treatment conditions.

For a slightly more complex illustration of the use of Eq. (11-4), consider the data presented in Table 10-1, in which the example from Chap. 6 was expanded to include a history lecture as a third condition. Suppose we were interested in conducting all three comparisons between pairs of means. As you know, the coefficients for these pairwise comparisons are $+1$ and -1 for the two critical means and 0 for all others (since they do not enter into the comparison). The three differences are:

Physical science versus social science

$$\begin{aligned}\hat{\psi}_1 &= (+1)(40.00) + (-1)(26.00) + (0)(34.50)\\ &= 40.00 - 26.00 = 14.00\end{aligned}$$

Physical science versus history

$$\begin{aligned}\hat{\psi}_2 &= (+1)(40.00) + (0)(26.00) + (-1)(34.50)\\ &= 40.00 - 34.50 = 5.50\end{aligned}$$

Social science versus history

$$\begin{aligned}\hat{\psi}_3 &= (0)(40.00) + (+1)(26.00) + (-1)(34.50)\\ &= 26.00 - 34.50 = -8.50\end{aligned}$$

We can now substitute the necessary values in Eq. (11-4) to find the sums of squares associated with these differences; that is,

$$SS_{comp.\ 1} = \frac{(12)(14.00)^2}{(+1)^2 + (-1)^2 + (0)^2} = \frac{2352.00}{2} = 1176.00$$

$$SS_{comp.\ 2} = \frac{(12)(5.50)^2}{(+1)^2 + (0)^2 + (-1)^2} = \frac{363.00}{2} = 181.50$$

$$SS_{comp.\ 3} = \frac{(12)(-8.50)^2}{(0)^2 + (+1)^2 + (-1)^2} = \frac{867.00}{2} = 433.50$$

For an example of a complex comparison, consider the contrast between the group receiving the history lecture and the two combined groups receiving the different science lectures. The standard set of coefficients for this comparison consists of $+\frac{1}{2}$, $+\frac{1}{2}$, and -1. The difference between the two means is

$$\begin{aligned}\hat{\psi}_4 &= (+\tfrac{1}{2})(40.00) + (+\tfrac{1}{2})(26.00) + (-1)(34.50)\\ &= 20.00 + 13.00 - 34.50 = -1.50\end{aligned}$$

Substituting in Eq. (11-4), we find

$$SS_{comp.\ 4} = \frac{(12)(-1.50)^2}{(+\frac{1}{2})^2 + (+\frac{1}{2})^2 + (-1)^2} = \frac{27.00}{1.50} = 18.00$$

Earlier we indicated that fractional coefficients may be transformed into whole numbers to simplify calculations. If we multiply the standard set of coefficients in this example by 2, we obtain $+1$, $+1$, and -2. With these new coefficients,

$$\begin{aligned}\hat{\psi} &= (+1)(40.00) + (+1)(26.00) + (-2)(34.50)\\ &= 40.00 + 26.00 - 69.00 = -3.00\end{aligned}$$

Substituting in Eq. (11-4), we have

$$SS_{comp.\ 4} = \frac{(12)(-3.00)^2}{(+1)^2 + (+1)^2 + (-2)^2} = \frac{108.00}{6} = 18.00$$

You should note that while these transformed coefficients do produce the same value for the sum of squares (18.00), the value of $\hat{\psi}$ (-3.00) does *not* represent the actual difference between the two means (-1.50). As you have seen, however, Eq. (11-4) compensates for this change. Thus, you can calculate $SS_{comp.}$ with any convenient set of coefficients that are derived from the standard set representing the difference under consideration.

Final Steps

The final steps consist of calculating a mean square, forming an *F* ratio, and evaluating a null hypothesis. A mean square is calculated by dividing a sum of squares by the appropriate number of *df*. Because the *df* associated with a difference between two means is 1—that is, $df_{comp.} = 1$—each comparison sum of squares is also the mean square. More explicitly,

$$MS_{comp.} = \frac{SS_{comp.}}{df_{comp.}} = \frac{SS_{comp.}}{1} = SS_{comp.}$$

An *F* ratio is formed by dividing each comparison mean square by the error term from the *omnibus* or *overall analysis*—that is, $MS_{S/A}$. More specifically,

$$F_{comp.} = \frac{MS_{comp.}}{MS_{S/A}} \qquad (11\text{-}5)$$

The statistical hypotheses for a single-*df* comparison may be expressed as follows:

$$H_0:\ \psi = 0$$
$$H_1:\ \psi \neq 0$$

where ψ represents the difference expressed in terms of population means. Rejecting the null hypothesis leads to the conclusion that the observed difference between two means $(\hat{\psi})$ is significant. The obtained value of *F* is evaluated in the usual manner, with $F_{comp.}$ being compared with the critical value of *F* listed under $df_{num.} = 1$, $df_{denom.} = df_{S/A}$, and α equal to the level you have chosen for planned comparisons (probably $\alpha = .05$).

You should note two important features of the *F* test: (1) the test evaluates precisely those comparisons you have earmarked for analysis during the planning

Table 11-1
Summary of Comparisons

Source	*SS*	*df*	*MS*	*F*
Comp. 1 (physical science vs. social science)	1176.00	1	1176.00	8.44*
Comp. 2 (physical science vs. history)	181.50	1	181.50	1.30
Comp. 3 (social science vs. history)	433.50	1	433.50	3.11
Comp. 4 (combined sciences vs. history)	18.00	1	18.00	.13
S/A		33	139.36	

* $p < .05$.

stages of the study, and (2) the *df* for the error term ($df_{denom.}$) are determined from the omnibus, *overall analysis* ($MS_{S/A}$) and *not* from the within-groups *df* associated with the particular groups involved in the analysis. That is, though the comparison may be contrasting conditions 1 and 2, the error term is based on information gathered from all the conditions in the experiment. The operation of both of these features means that Eq. (11-5) provides a powerful test of hypotheses generated by planned comparisons.

The analyses of the four comparisons we considered in the last section are summarized and completed in Table 11-1. Each *F* is formed by dividing the comparison mean square by the error term obtained from the omnibus analysis ($MS_{S/A} = 139.36$). The critical value of *F*, which is based on $df_{num.} = 1$ and $df_{denom.} = 33$, is approximately 4.17 at $\alpha = .05$.[2] Only the comparison between the two science lectures (physical science versus social science) is significant.

Orthogonal Comparisons

Central to an understanding of ANOVA and MRC is the fact that a sum of squares can be subdivided into separate and independent components. In ANOVA, we have seen that SS_T may be divided into two useful components, namely, the between-groups sum of squares SS_A and the within-groups sum of squares $SS_{S/A}$. In MRC, the equivalent breakdown consists of dividing the total variability in the dependent variable (SS_Y) into two sums of squares, one representing the variability associated with group membership ($SS_{reg.}$) and the other a residual sum of squares—the variability in *Y* (the dependent variable) not accounted for by group membership ($SS_{res.}$).

[2] We have used $F(1, 30)$ for the critical value, which is a slightly larger value than required and yields an actual significance level that is somewhat smaller than .05.

It is generally the case that *any* sum of squares with more than 1 *df* can be divided into two or more independent sums of squares, and that the maximum number of such subdivisions is equal to the number of degrees of freedom associated with the sum of squares being subdivided. For the overall analysis, however, we are interested only in the between-groups and within-groups sums of squares. For planned comparisons (or single-*df* comparisons conducted following the omnibus test), we are obviously interested in what we might call *comparison* sums of squares—usually single-*df* comparisons—that contribute to the overall between-groups variability.

The statement that any sum of squares can be divided into two or more independent sums of squares holds only for what are known as **orthogonal comparisons**. Two single-*df* comparisons are said to be orthogonal if they reflect ***independent*** or completely ***nonoverlapping*** pieces of information. What this means is that the outcome of one comparison gives no indication whatsoever about the outcome of the other comparison. If all comparisons are orthogonal to one another, we refer to them as a set of **mutually orthogonal comparisons**. Thus, we can say that the SS_A can be broken down into a set of $df_A = a - 1$ mutually orthogonal single-*df* comparisons. In symbols,

$$SS_A = SS_{comp.\,1} + SS_{comp.\,2} + \cdots + SS_{comp.\,a-1} \qquad (11\text{-}6)$$

where the $a - 1$ comparisons are mutually orthogonal.

The orthogonality of any two single-*df* comparisons is easily determined by comparing the coefficients defining the two comparisons. Let us call the coefficients for one comparison c_i and the coefficients for the other comparison c'_i. The two comparisons are orthogonal if

$$(c_1)(c'_1) + (c_2)(c'_2) + (c_3)(c'_3) + \cdots = 0$$

or, more compactly,

$$\Sigma\,(c_i c'_i) = 0 \qquad (11\text{-}7)$$

To illustrate, suppose in our three-condition experiment we were concerned with a comparison of a_1 and a_3; the coefficient would be $+1$ for a_1, 0 for a_2, and -1 for a_3. Now suppose that another comparison of interest was the average of a_1 and a_3 versus a_2; here the coefficients would be $+\frac{1}{2}$ for a_1, -1 for a_2, and $+\frac{1}{2}$ for a_3. Application of Eq. (11-7) yields the following:

$$(+1)(+\tfrac{1}{2}) + (0)(-1) + (-1)(+\tfrac{1}{2}) = \tfrac{1}{2} + 0 - \tfrac{1}{2} = 0$$

Given this result, our two comparisons are orthogonal and are not providing redundant information. Thus, orthogonality is verified simply by multiplying corresponding pairs of coefficients—one pair for each level of factor *A*—and determining

Table 11-2
Coefficients for Single-*df* Comparisons

	Physical Science	Social Science	History
Comparison 1	+1	−1	0
Comparison 2	+1	0	−1
Comparison 3	0	+1	−1
Comparison 4	$+\frac{1}{2}$	$+\frac{1}{2}$	−1

that the sum of the products equals zero. Any value other than zero indicates that the two comparisons are *not* orthogonal.

In Table 11-1, we tested the significance of the difference between each pair of group means, as well as the difference between the mean for the two combined science groups and the mean for the history group. The coefficients for these comparisons are presented in Table 11-2. From Eq. (11-6), we know that the SS_A can be subdivided into a total of two orthogonal comparisons ($a - 1 = 3 - 1 = 2$). As surprising as it may be, only *one* set of orthogonal comparisons can be formed with the comparisons in Table 11-2. What about the first two comparisons? If we substitute the coefficients for these two comparisons in Eq. (11-7), we find

$$(+1)(+1) + (-1)(0) + (0)(-1) = 1 + 0 + 0 = 1$$

Since the sum is not zero, the two comparisons are not orthogonal. In fact, no set containing any two of the pairwise comparisons is orthogonal. (You may wish to verify this statement for yourself.) The only remaining possibility is a set that contains the fourth comparison and one of the pairwise comparisons. Applying the test to comparisons 1 and 4, we find that this is the orthogonal set:

$$(+1)(+\tfrac{1}{2}) + (-1)(+\tfrac{1}{2}) + (0)(-1) = \tfrac{1}{2} - \tfrac{1}{2} + 0 = 0$$

Earlier in this section, we indicated that a complete set of mutually orthogonal comparisons fully accounts for the original sum of squares. We can illustrate this property by adding together the sums of squares for comparisons 1 and 4 to show that they sum to SS_A. From Table 11-1, we see that

$$SS_{comp.\,1} + SS_{comp.\,4} = 1176.00 + 18.00 = 1194.00$$

From our earlier calculations (see Table 10-2). we find that this sum is exactly equal to the SS_A. What about other sets of comparisons? For comparisons 1 and 2, for example, the sum is

$$SS_{comp.\,1} + SS_{comp.\,2} = 1176.00 + 181.50 = 1357.50$$

which does not equal SS_A. If you try any other set of comparisons, you will find that none sums exactly to the value of SS_A (1194.00).

What implications does orthogonality have for a researcher? Some authorities suggest that all planned comparisons should be mutually orthogonal. One reason commonly given for this recommendation is that orthogonal comparisons represent an efficient use of an experimental design—a division of between-groups variability into a tidy set of nonoverlapping sources of variability. Most authors of statistics texts for behavioral scientists disagree with this recommendation, however. They argue that orthogonality should not be a *requirement* of planned comparisons. Instead, they feel that the overriding considerations in selecting a set of planned comparisons are the following:

1. **The set should be an integral part of the experimental design.**
2. **The comparisons should represent the primary purpose of the experiment.**
3. **The comparisons should constitute meaningful and direct tests of the research hypotheses.**

The entire set of comparisons we conducted on the data from the vocabulary experiment (see Table 11-1) seems to satisfy these criteria, even though the comparisons are not mutually orthogonal. *All four* comparisons provide useful information concerning the outcome of the experiment.

11.3 SINGLE-*df* COMPARISONS: THE MRC APPROACH

You may recall that in Chap. 10 we demonstrated the equivalency of MRC and ANOVA by using *contrast coding* to represent the levels of the independent variable and by using the R^2 to evaluate the omnibus null hypothesis. The purpose of this section is to show how contrast-coded vectors can be used to conduct the same sorts of single-*df* comparisons we considered with ANOVA in Sec. 11.2.

Contrast Coding

Contrast coding, you will remember, involves the assignment of values to the subjects so that different comparisons between groups are represented by each vector. For the overall analysis, the coding principle requires us to establish $a - 1$, or two, vectors in our current example, to permit the calculation of the omnibus R^2.

In Chap. 8 we introduced decision rules that we could use to fix the probability of a type I error at α for the *omnibus* test. If we apply these same rules to single-*df* comparisons (like the comparisons that were the subject of Chap. 11), we in effect fix the type I error at α for *each one of the statistical tests* conducted in the analysis of the experiment. The consequences of applying the rules to several tests at once will be our concern in this chapter.

12.1 PROBLEMS ASSOCIATED WITH ANALYTICAL COMPARISONS

A serious problem exists whenever we perform more than one statistical test in the analysis of any research or experiment: each comparison—whether planned or post hoc—increases our chances of committing a type I error *somewhere* within the entire analysis. In order to talk about this problem, it is convenient to introduce two new terms, *per comparison* and *familywise* type I errors. We will now consider how these two ways of conceptualizing type I error are related.

Per comparison (*PC*) type I error is the type I error associated with the significance level that we set for any given statistical test. In most cases, a researcher would set the significance level at $\alpha_{PC} = .05$ for all comparisons. What effect does this decision have on type I error? If our focus is on the level of the individual statistical test, we can say that the type I error is .05 for each one of the tests. But suppose we consider a different point of reference, namely, the type I error for the *experiment as a whole*, which includes the entire set of comparisons tested in the analysis. If we do this, the separate per comparison probabilities actually combine to produce a much larger value, which we will call the **familywise (*FW*) type I error**. This category of error, which has also been called the **experimentwise type I error**, refers to the probability (α_{FW}) that at least one type I error has been committed *somewhere among* the various tests conducted in the analysis. If two tests are conducted, for example, the familywise error will approximately equal the sum of the two *PC* probabilities, namely, .10 (.05 + .05). If there are three tests, α_{FW} will approximately equal .15 (.05 + .05 + .05).

The exact relationship between α_{FW} and the number of statistical tests can be determined by the following formula:

$$\alpha_{FW} = 1 - (1 - \alpha_{PC})^c \qquad (12\text{-}1)$$

where *c* represents the number of *orthogonal* comparisons that are conducted. With the *PC* type I error set $\alpha_{PC} = .05$ and with $c = 3$ orthogonal comparisons, for example,

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

The same basic relationship between *FW* error, on the one hand, and the number of tests and *PC* error, on the other, holds for nonorthogonal comparisons as well, although the relationship is more complex.

Although researchers have known about the relationship expressed in Eq. (12-1) for some time, they still do not agree on what should be done about it. In reality, each researcher must decide (and justify to others) the steps taken to control *FW* error. What we hope you will extract from the present discussion is an appreciation for the nature of the problem and an understanding of some of the solutions that have been offered. This will better prepare you for determining your own response to the problem of *FW* error.[1]

Possible solutions to the problem of controlling *FW* error are many, but all reduce to the same mechanism, namely, *a decrease in the size of the rejection region* used to evaluate the significance of comparisons. Let us see how this general procedure works. As an example, suppose we planned to conduct five orthogonal comparisons. We know from Eq. (12-1) that if we used $\alpha_{PC} = .05$, the *FW* error would be

$$\alpha_{FW} = 1 - (1 - .05)^5 = 1 - (.95)^5 = 1 - .774 = .226$$

A relatively simple way of reducing *FW* error in this example would be to use a smaller probability for α_{PC}—that is, a higher level of significance for evaluating the *individual comparisons*. More specifically, consider what would happen to *FW* if we use $\alpha_{PC} = .01$ rather than .05 to assess the significance of each of these five comparisons. Turning again to Eq. (12-1), we find the new familywise type I error to be

$$\alpha_{FW} = 1 - (1 - .01)^5 = 1 - (.99)^5 = 1 - .951 = .049$$

which, as you can see, apparently solves the problem of increased *FW* simply and neatly! That is, familywise error is now equivalent to the significance level adopted by most researchers for omnibus statistical tests.

Before you become too complacent with this solution to the problem, you should realize that this control of *FW* error has been accomplished by increasing the probability of another kind of error, namely, *type II error*. This "cost" for controlling *FW* error can also be expressed as a *loss of power*. Let us consider this important point in more detail.

If we use the .01 level of significance rather than the .05 level to evaluate comparison null hypotheses, we are able to reduce familywise error, as you have seen. But as you can also see, this reduction is accomplished by the simple expedient of rejecting *fewer null hypotheses*. We control *FW* error by making it more

[1] You will find detailed discussions of the general problem in most advanced statistics texts. For an elaboration of the views set forth in this section, see Keppel (1982, Chap. 8).

difficult to reject null hypotheses; the fewer null hypotheses we reject, the lower our *FW* error. This is exactly what we want to do when the null hypothesis is *true*, of course. But the null hypothesis may also be *false*: some comparisons reflect real differences in the population. Requiring a higher level of significance for these comparisons directly increases type II error, and by definition, decreases power. The problem, then, is to find a way of *balancing* the two kinds of errors.

12.2 PLANNED COMPARISONS

Most recommendations concerning the control of familywise type I error distinguish between planned and post hoc comparisons. Recommendations for planned comparisons usually do not include a correction for familywise error, except perhaps, when the number of the planned comparisons exceeds some reasonable value such as the degrees of freedom for the treatment sum of squares (see Keppel, 1982, pp. 147–150). This disregard for *FW* error is generally defended by the argument that planned comparisons typically constitute the primary purpose of a study, and as such, they should be subjected to the most sensitive statistical test possible. This type of test is one that treats each comparison as if the experiment were specifically designed to focus on it. Any increase in *FW* error resulting from the statistical assessment of planned comparisons is thus accepted as one of the calculated risks of experimentation.

12.3 POST HOC COMPARISONS

Most corrections for familywise error are applied to comparisons conducted after the data have been initially examined and analyzed. Post hoc tests are treated differently from planned comparisons because of their potentially large number and because of their fortuitous, unplanned nature. When one is sifting through a set of data in search of significant differences, considerably more comparisons are examined and assessed than are ever proposed in the planning stage of an experiment. Familywise type I error under these circumstances can be intolerably high.

Recommendations for controlling *FW* error for post hoc comparisons depend on the nature of the pool of differences being examined. We will consider three common situations, in which the pool consists of (1) all possible comparisons, (2) all possible differences between pairs of treatment means, and (3) all possible differences between a control condition and a number of experimental or treatment conditions. The procedures are applicable to all single-factor experiments, regardless of whether the *F* tests are conducted with ANOVA or MRC. The logic is the same.

All Possible Comparisons

When complex comparisons are included in the total pool of potential comparisons—all the comparisons that one might examine when combing through data—the post hoc pool is often very large indeed, and a severe correction is usually required to keep *FW* error at a reasonable level. The **Scheffé test** was designed for exactly this type of situation (Scheffé, 1953). The test is simple to perform. All we do is calculate a new critical value of F (F_S) to incorporate into the decision rules, as follows:

$$F_S = (a - 1)F(df_A, df_{S/A}) \qquad (12\text{-}2)$$

where a equals the number of treatment conditions and $F(df_A, df_{S/A})$ is the critical value of F for the *omnibus* F test.[2] This value is found in Table A-1 with $df_{num.} = a - 1$ and $df_{denom.} = df_{S/A}$. (Please note that $df_{num.}$ does *not* equal $df_{comp.}$; the two are often mistakenly equated in applications of the Scheffé test.) One's choice of α at this point determines the maximum value that α_{FW} will ever reach regardless of the number of comparisons actually evaluated. It is this property that makes the Scheffé test particularly attractive to researchers who are poring over large data sets, searching for significant differences.

As an example of the calculations, consider again the four comparisons we analyzed as planned comparisons in Chap. 11. We will treat them now as post hoc comparisons and evaluate them by using the Scheffé test. For Eq. (12-2), we need a ($a = 3$, in this example) and the critical value for the omnibus F test, $F(2, 33) = 3.32$ at $\alpha = .05$. Substituting in the formula, we find

$$F_S = (3 - 1)(3.32) = (2)(3.32) = 6.64$$

We would now use 6.64 as the critical value of F to test the significance of these (and any other) comparisons we conduct, whether with ANOVA or with MRC.

As we have noted already, the Scheffé test guarantees that the *FW* error will be no greater than the value of α used to enter the F table (.05 in this case), no matter how many comparisons are conducted. An inspection of the $F_{comp.}$'s in either Table 11-1 or Table 11-4 indicates that the comparison between the two

[2] In MRC terminology, $df_A = k$ (the number of vectors required for the omnibus R^2) and $df_{S/A} = N - k - 1$ (the degrees of freedom for the residual sum of squares).

science conditions would still be significant under the Scheffé test. You should realize, of course, that only the largest differences will emerge triumphant from an application of the Scheffé test. One way to make this point is to compare the value of F_S (6.64) with the value of the uncorrected F we would use for planned comparisons ($F = 4.17$). Comparisons producing $F_{comp.}$'s that fall between these two critical values would be declared significant if they were planned comparisons and not significant if they were post hoc comparisons subjected to the Scheffé correction.

In summary, the Scheffé test provides protection from *FW* type I error when a researcher hopes to discover interesting, but still unexpected, differences between treatment conditions and combinations of treatment conditions. Since the total pool of such comparisons is relatively large, so must be the correction required to restore the *FW* rate to reasonable levels. As we have already noted, however, the "cost" of this protection is a considerable loss in the power to detect real treatment differences. This loss of power may be substantially reduced if something can be done to restrict the size of the comparison pool examined by a researcher. A smaller pool requires a smaller reduction in the α_{PC} to exercise the desired control over *FW* error. We will next consider two procedures that capitalize on this strategy of restricting one's attention to certain smaller and better-defined subsets of comparisons.

All Possible Differences between Pairs of Means

One obvious way of reducing the pool of post hoc comparisons is to concentrate on the differences between *pairs of treatment means* and simply not consider complex comparisons in the post hoc analysis. To see how this reduces the number of comparisons, let us consider several examples: if $a = 3$, for instance, the total pool contains a combination of 6 pairwise and complex comparisons, while the smaller pool consists of 3 pairwise comparisons; if $a = 4$, the total pool contains 25 comparisons, while the smaller pool contains 6; and finally, if $a = 5$, the total pool contains 90 comparisons, while the smaller pool contains 10. Because of the difference in the size of these two pools, α_{PC} requires a smaller adjustment for pairwise comparisons than it requires when the pool contains both pairwise and complex comparisons. One test that provides control over the smaller pool of differences is called the **Tukey test** (Tukey, 1953).

The Tukey test is most easily performed by calculating the differences between all pairs of means and comparing them against a *minimum*, or *critical*, difference that must be exceeded for an observed difference to be declared significant. This critical value $\bar{d}_T$ is given by the formula

$$\bar{d}_T = \frac{q_T\sqrt{MS_{S/A}}}{\sqrt{s}} \qquad (12\text{-}3)$$

where q_T = an entry in a special table, called the **studentized range statistic** (Table A-2)

$MS_{S/A}$ = the error term from the overall analysis of variance

s = the sample size of the treatment groups

Using the data from our example to illustrate the Tukey test, we first find the three differences, namely,

$$\bar{Y}_{A_1} - \bar{Y}_{A_2} = 40.00 - 26.00 = 14.00$$
$$\bar{Y}_{A_1} - \bar{Y}_{A_3} = 40.00 - 34.50 = 5.50$$
$$\bar{Y}_{A_2} - \bar{Y}_{A_3} = 26.00 - 34.50 = -8.50$$

Next, we calculate the critical difference $\bar{d}_T$ by using Eq. (12-3). The value for q_T, which is required by the formula, is found by entering Table A-2 and coordinating three quantities, df_{error} (the *df* associated with the $MS_{S/A}$), k (the number of treatment means—a in this design), and α_{FW} (the *FW* error rate chosen for the Tukey test). Using $df_{error} = 30$ (since $df_{error} = 33$ does not appear in the table), $k = 3$, and $\alpha_{FW} = .05$, we find $q_T = 3.49$. Substituting this and the other required values ($MS_{S/A} = 139.36$ and $s = 12$) in Eq. (12-3), we find

$$\bar{d}_T = \frac{(3.49)\sqrt{139.36}}{\sqrt{12}} = \frac{41.22}{3.46} = 11.91$$

A comparison of the three observed differences against the new criterion, $\bar{d}_T = 11.91$, indicates that only the difference between the physical science and social science groups (14.00) is significant.

Under some circumstances, you may wish to conduct the Tukey test in the same way we performed the Scheffé test, namely, in conjunction with the F statistic. This method is ideally suited for analysis by MRC, where the computer output for these comparisons is expressed in terms of zero-order correlations from which F ratios can be formed. The new critical value of F required by the Tukey test (F_T) is given by

$$F_T = \frac{(q_T)^2}{2} \qquad (12\text{-}4)$$

In the present case,

$$F_T = \frac{3.49^2}{2} = 6.09$$

The Tukey test is conducted by using $F_T = 6.09$ as the criterion for evaluating the $F_{comp.}$'s. An inspection of Tables 11-1 and 11-4, where these statistical analyses are summarized, permits the same conclusion, namely, that only the difference between physical science and social science is significant.

It is instructive to compare the critical values of F for the Scheffé and the Tukey tests. With this example, the critical value of F was slightly greater for the Scheffé test (6.64) than that required by the Tukey test (6.09). The difference between these two critical values increases as the number of treatment conditions in an experiment increases.

All Possible Comparisons between a Control and Several Treatment Conditions

A final type of situation involves an even smaller pool of potential comparisons—an experiment in which one condition, usually a control or baseline condition of some sort, is compared against a number of experimental or treatment conditions. As you would suspect, the degree to which this restricted comparison pool is smaller than the other pools increases directly with the scope of the experiment. To illustrate, if $a = 3$, then 2 of the 3 pairwise comparisons represent differences between the control and the two experimental conditions; if $a = 4$, the numbers are 3 out of 6; and if $a = 5$, they are 4 out of 10. Because the comparisons involved are fewer than those considered by either the Scheffé test or the Tukey test, the correction for FW error will not be as severe as that given by either of those tests. The test developed for this type of situation is known as the **Dunnett test** (Dunnett, 1955).

Like the Tukey test, the Dunnett test is most simply conducted by comparing the differences between the control and experimental means against a critical difference that must be exceeded to be significant at the chosen α_{FW} level. The formula for calculating this difference ($\bar{d}_D$) is

$$\bar{d}_D = \frac{q_D\sqrt{2\,MS_{S/A}}}{\sqrt{s}} \qquad \text{(12-5)}$$

where q_D is an entry in Table A-3 of Appendix A and the other quantities are familiar to you. The value of q_D is determined by the total number of conditions (k) involved in the analysis, the degrees of freedom associated with the error term ($df_{S/A}$), and the value chosen for FW error (α_{FW}).[3]

[3] The values of q_D given in the first part of Table A-3 are for situations in which researchers are interested in the possibility of positive as well as negative differences between the control and experimental conditions (called *Two-Tailed Comparisons*). A special table is available for the less common situation in which researchers are concerned only with differences in "one direction"—that is, either positive or negative differences, not both. This other table is included in Table A-3 as *One-Tailed Comparisons.*

We will calculate $\bar{d}_D$ using the example from the last section. To find q_D, we look for the entry in Table A-3 at $k = 3$, $df_{error} = 30$, and $\alpha_{FW} = .05$. This value is 2.32. Substituting in Eq. (12-5), we find

$$\bar{d}_D = \frac{(2.32)\sqrt{(2)\,(139.36)}}{\sqrt{12}} = \frac{38.72}{3.46} = 11.19.$$

As expected, this critical difference is slightly smaller than that required for the Tukey test performed on the same data ($\bar{d}_T = 11.91$).

If you wish to work with the F test, you can use

$$F_D = (q_D)^2 \qquad \text{(12-6)}$$

as the critical value with which to evaluate $F_{comp.}$. In the present case,

$$F_D = 2.32^2 = 5.38$$

This critical value can be compared with corresponding values for the Scheffé test (6.64) and the Tukey test (6.09) to illustrate the different sensitivities of the three tests.

12.4 THE FISHER TEST

An entirely different approach to the problem of FW type I error is the **protected least significant difference test**, which we will call the **Fisher test** (Fisher, 1949). The test centers on the outcome of the *omnibus* F test; the significance or nonsignificance of this test determines whether additional tests will be conducted at all. If the F is significant, comparisons are evaluated *without* correction for FW error; if the F is not significant, no further tests are conducted.

The Fisher test is most appropriate in situations in which initially, at least, all treatments are given equal consideration, that is, there are no favored treatments or anticipated outcomes—in short, where there are no planned comparisons as we have defined them. An example would be an experiment comparing consumer preferences among alternative ways of packaging a certain product. The object of the study is to discover whether the different package designs make any difference to the potential consumer. This is where the omnibus F test comes into play: it assesses the average differences associated with the treatment conditions. Only if the overall null hypothesis is rejected does the investigator examine the specific differences between conditions to find out which are the best and which are the worst.

Familywise type I error is controlled *indirectly* by the Fisher test. The omnibus *F* test acts as a "filter," which permits additional tests only when the evidence looks "good"—that is, when treatment differences are sufficiently large not to be reasonably attributed to the operation of chance factors. Although it is true that a researcher will sometimes falsely conclude that differences are present in the population when in fact they are not, this does not happen very often (only 5 percent of the time when $\alpha = .05$). Thus, little long-term risk is incurred by following the Fisher procedure. The Fisher test has been studied by statisticians and shown to offer an excellent balancing of type I and type II errors (see Carmer & Swanson, 1973). We do not suggest its use with studies in which planned comparisons are also present, for reasons that have been expressed elsewhere (Keppel, 1982, pp. 158–159).

12.5 RECOMMENDATIONS AND COMMENT

We recommend that planned comparisons be evaluated without undue concern for their effect on familywise type I error. Furthermore, we recommend that planned comparisons be the strategy for research. However, if there is no reason or it is not feasible to conduct planned comparisons, then post hoc comparisons are the alternative. But most researchers become concerned at this point with the greatly increased potential for familywise error associated with post hoc comparisons and adopt some strategy for dealing with it. We have described three techniques which have been developed to correct *FW* error for different pools of possible comparisons. If a mixture of complex comparisons and differences between pairs of means are candidates for post hoc tests, we recommend the Scheffé test. On the other hand, if only pairwise differences are of interest, we recommend the Tukey test. We recommend the Dunnett test when only differences between a control and experimental conditions are involved. The Fisher test seems most appropriate for situations in which planned comparisons are not an integral part of the experimental design.

These recommendations are summarized as a flowchart in Fig. 12-1. We begin at the top with the question "Do you have planned comparisons?" Your answer to this question branches to additional questions and finally to the appropriate set of procedures. If you answer yes, for example, you perform the planned comparisons without undertaking any correction for *FW* error; if you wish to follow these with additional post hoc comparisons, you would select one of the correction procedures we have discussed—Scheffé, Tukey, or Dunnett—depending on the nature of the comparisons you have selected to examine. On the other hand, if you answer no to the initial question, you may perform either the Fisher test or one of the three alternative post hoc tests. These possibilities are presented at the bottom of the flowchart.

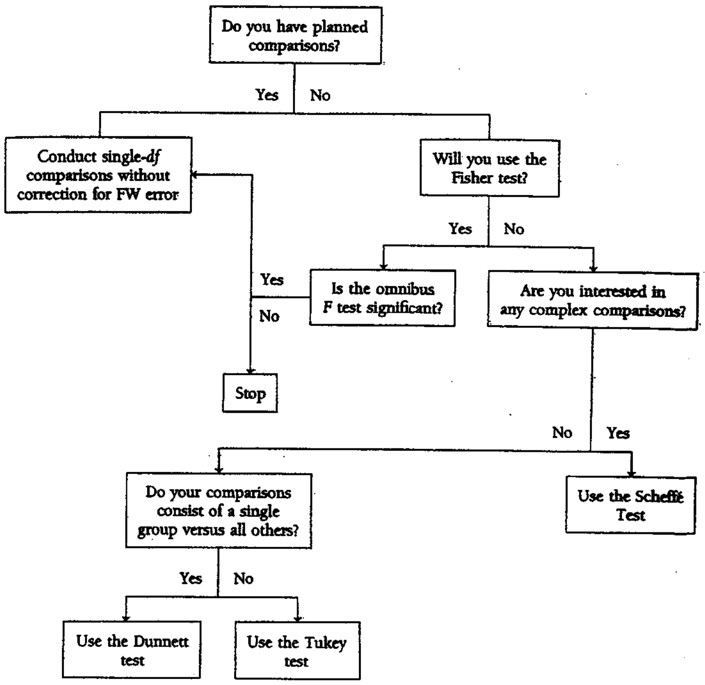

Do you have planned comparisons?
Yes
No
Conduct single-*df* comparisons without correction for FW error
Will you use the Fisher test?
Yes
No
Yes
No
Is the omnibus F test significant?
Are you interested in any complex comparisons?
Stop
No
Yes
Do your comparisons consist of a single group versus all others?
Use the Scheffé Test
Yes
No
Use the Dunnett test
Use the Tukey test

Figure 12-1 Schematic representation of post hoc techniques.

Any correction for *FW* error disregards another important concern of researchers, namely, type II error, or the loss of power created whenever an *FW* correction is incorporated into the evaluation process. A constructive suggestion for dealing with this problem is to expand the usual decision rules in which we either reject or do not reject a null hypothesis to include a third course of action, the opportunity to *suspend judgment*. That is, suppose we decided to reject the null hypothesis only when the F_{comp}. exceeds the *familywise* criterion but to make *no* *decision*—that is, to suspend judgment—when an F_{comp}. happens to fall between the criterion for ***planned comparisons*** and the familywise criterion. To be more explicit, consider the following modified decisions rules:

If $F_{observed} \geq F_{FW}$, reject H_0,
If $F_{observed} < F_{PC}$, do not reject H_0,
If $F_{observed}$ falls between F_{PC} and F_{FW}, *suspend judgment.*

The first of the decision rules concerns the familywise criterion; an $F_{observed}$ that exceeds this critical value (F_{FW}) will be rejected without question. The second rule concerns the "normal" per comparison criterion applied to uncorrected planned comparisons; an $F_{observed}$ that falls short of this critical value (F_{PC}) will not be rejected. The final rule pertains to our decision when the $F_{observed}$ falls between these two critical values—we suspend judgment. We apply this third rule when we come across an unexpected finding which, if it were the result of a planned comparison, we would have termed significant; we suspend judgment rather than commit a promising finding to potential obscurity by labeling it "not significant" under the more stringent *FW* criterion.[4]

12.6 SUMMARY

One consequence of assessing a number of single-*df* comparisons is an increase in the probability of committing type I errors during the course of the entire analysis. This probability, which is known as *familywise type I error*, increases directly with the number of statistical tests performed, whether they are planned tests or not. Researchers tend to ignore familywise error when doing planned comparisons, however, but usually adopt some way of reducing it with unplanned comparisons. We discussed several techniques that can be used with different types of comparisons. The Scheffé test is used when the pool of possible single-*df* comparisons consists of all pairwise and complex comparisons. The Tukey test is used when the pool consists of all pairwise comparisons. The Dunnett test is used when a single control condition is compared with a number of treatment conditions. The Fisher test, which focuses on the outcome of the omnibus *F* test, is recommended for situations in which there are no planned comparisons motivating the research. Finally, we reiterate that the above concerns are related to the number of comparisons undertaken and *not* to whether the analytical strategy one adopts is ANOVA or MRC.

[4] This suggestion is developed more fully in Keppel (1982, pp. 162–164).

12.7 EXERCISES

1. The experiment described in Problem 1 of Chap. 10 consisted of three conditions, praise for correct responses ("praise"), reproof for mistakes ("reproof"), and no verbal comment at all ("none"). For the questions below, assume that you had no planned comparisons when you designed the experiment. What procedure for controlling *FW* error (Dunnett, Scheffé, or Tukey) is most appropriate under the following circumstances?

a. All possible comparisons between pairs of means
b. Comparisons between each of the two "verbal" conditions and the condition receiving no verbal comment
c. A comparison between "none" and the combined conditions receiving verbal comment of some sort and a comparison of the two verbal conditions

2. Using the data from Problem 1 of Chap. 10,

a. Conduct a Tukey test.
b. Conduct a Dunnett test.
c. What is the critical value for a Scheffé test?

You have seen how a single-factor experiment can be analyzed to yield information concerning a number of research hypotheses. If the manipulation is qualitative in nature, the analysis generally takes the form of assessing miniature two-group "experiments." On the other hand, if the manipulation is quantitative in nature, the analysis usually consists of the examination and assessment of trend components presumed to underlie the relationship between the independent and dependent variables. (We consider trend analysis in Chap. 23.) In either case, the experimental manipulation is conceived as a *single* independent variable—a variation either in the type or in the amount of the independent variable.

We will examine a different kind of design in this chapter, one in which two independent variables (factor *A* and factor *B*) are manipulated *simultaneously* within the context of the same experiment. This type of design, known as the **factorial design**, is quite common in the behavioral sciences, for the important reason that it greatly expands the sorts of questions one can study in an experiment. In this chapter, we will consider the nature of the overall analysis of a two-factor design and show how the analysis is accomplished by ANOVA and MRC. In Chaps. 14 and 15, we will turn our attention to the detailed analysis of the factorial experiment.

13.1 THE OVERALL ANALYSIS

Suppose our experiment comparing the relative merits of teaching vocabulary words in the context of three different types of lectures (factor *A*) is expanded to include a second independent variable, mode of presentation (factor *B*). More specifically, suppose we are comparing computer-assisted instruction with a "standard" method of presentation (a lecture given by a teacher). A factorial design combines the two independent variables in such a way that all possible combinations of the levels of the two variables are represented in the experiment. In the present case, this would mean that different groups of students would receive lectures on physical science, social science, and history, under each of the two methods of presentation, computer or standard, and thus there would be a total of 3 × 2 = 6 treatment conditions.

This arrangement is diagramed in Table 13-1, where you can see the exact nature of the design. Each cell represents a different treatment condition of the experiment formed by a unique combination of the levels of the two independent variables. In a completely randomized factorial experiment, which is the kind of experiment we are considering in this chapter, subjects are assigned randomly to the different treatment conditions. Typically, each group is represented by an equal number of subjects (s). (Factorial experiments with unequal sample sizes are discussed in Chap. 24.)

Table 13-1
An Example of a Two-Factor Design

Method of Presentation (Factor *B*)	Type of Lecture (Factor *A*)		
	Physical Science (a_1)	Social Science (a_2)	History (a_3)
Computer (b_1)			
Standard (b_2)			

Factorial experiments are usually described in terms of the number of levels associated with the two independent variables. The present example would be called a *completely randomized* 3 × 2 ("three-by-two") *factorial design*, which clearly specifies the fact that two independent variables have been manipulated factorially, one with three levels and the other with two levels, and that the total number of treatment conditions is six. By *completely randomized* we mean that individual subjects are randomly assigned to only one of the treatment combinations; in other types of designs subjects may each receive more than one treatment combination.

Let us now fill in the data for this example. In the upper portion of Table 13-2, we present vocabulary scores for all the sets of $s = 6$ subjects constituting the different treatment groups. (These data are derived from Table 10-1; the first six scores for each lesson have become the "computed-assisted" scores, and the remaining six scores are the "standard" scores.) The means of the groups are listed in the lower portion of the table in what we will call a *factorial matrix*.

You will notice that the two "margins" of the matrix contain the averages of the means in the individual columns (the column marginal means) and in the individual rows (the row marginal means). The marginal means can be thought of as the results of two "artificial" *single*-factor experiments, one in which the independent variable is the type of lecture (the column marginal means, 40.00, 26.00, and 34.50) and another in which the independent variable is the method of presentation (the row marginal means, 41.33 and 25.67). These average effects are artificial in the sense that they are not the product of an *actual* single-factor design in which only one independent variable is manipulated, but rather are the effects produced by averaging or "collapsing" over the levels of the other independent variable manipulated in the factorial experiment.

Table 13-2
Numerical Example

	Vocabulary Scores					
	Physical Science		Social Science		History	
	Computer	Standard	Computer	Standard	Computer	Standard
	53	44	47	13	45	46
	49	48	42	16	41	40
	47	35	39	16	38	29
	42	18	37	10	36	21
	51	32	42	11	35	30
	34	27	33	6	33	20
Sum:	276	204	240	72	228	186
Mean:	46.00	34.00	40.00	12.00	38.00	31.00

Matrix of Means

Method of Presentation (Factor B)	Type of Lecture (Factor A)			
	Physical Science (a_1)	Social Science (a_2)	History (a_3)	Average
Computer (b_1)	46.00	40.00	38.00	41.33
Standard (b_2)	34.00	12.00	31.00	25.67
Average	40.00	26.00	34.50	

These two sets of average effects are called the **main effects** of the two variables. The term *main effect* is not to be interpreted to mean "primary" or "important." Whether main effects are of any systematic interest to a researcher depends primarily on the *joint influence* of the two variables, which is revealed by an examination of the means within the body of the factorial matrix.

Consider, for example, the three means in the first row of the matrix (46.00, 40.00, and 38.00). These means reflect the effects of the three lectures for those students receiving the computer presentation. They show that the group receiving the lecture on physical science has surpassed the group receiving the lecture on social science by 6.00 words (46.00 − 40.00) and the group receiving the lecture on history by 8.00 words (46.00 − 38.00), and that social science has surpassed history by 2.00 words (40.00 − 38.00).

What about the corresponding manipulation under the standard presentation? The three means in the second row (34.00, 12.00, and 31.00) show a striking change in the effects of the three types of lecture, namely, a very substantially increased difference between physical science and social science (34.00 − 12.00 = 22.00 words) and a somewhat reduced difference between physical science and history (34.00 − 31.00 = 3.00 words). The difference between social science and history actually shows a reversal for the two methods of presentation, namely, a *marked inferiority* for social science with the standard presentation (a difference of 19.00 words) as opposed to the *small superiority* (2.00 words) found with the computer presentation. It appears, then, that the type of lecture produces different results with the two methods of presentation. It is for this reason, therefore, that the *average* effects of the type of lecture—reflected by the column marginal means—represent a distorted picture of the results of the actual factorial experiment revealed by the two sets of row means within the body of the matrix.

We reach a similar conclusion if we consider the difference between the two methods of presentation. While we find the computer method to be superior when we examine the row marginal means (41.33 − 25.67 = 15.66 words), the *magnitude* of the difference depends on the type of lecture. This is easily seen if we examine the pairs of row means column by column, within the body of the factorial matrix. More specifically, for physical science the difference is 12.00 words (46.00 − 34.00), for social science it is 28.00 words (40.00 − 12.00), and for history it is 7.00 words (38.00 − 31.00). Again, we see that the main effect of an independent variable is not representative of the results of the actual factorial experiment.

This type of situation—in which the effects of one of the independent variables depend on the levels of the other independent variable—is called **interaction.** Stated another way,

> **Interaction is present when the *pattern of differences* associated with either one of the independent variables changes as a function of the levels of the other independent variable.**

When this happens, the main effects do not yield a faithful picture of the results of a factorial experiment. You should note that this definition of interaction is technically correct only for population treatment means. In an actual experiment, which is assumed to be drawn randomly from these populations, the presence or absence of interaction is assessed by an *F* test designed for that purpose. We will discuss this statistical test shortly.

The way in which interactions operate can be seen by examining 2 × 2 designs. For example, suppose we have the following 2 × 2 layout with the cells labeled as shown:

$$\begin{array}{cc|c|c|} & & \multicolumn{2}{c}{A} \\ & & a_1 & a_2 \\ \cline{3-4} B & b_1 & \mu_1 & \mu_2 \\ \cline{3-4} & b_2 & \mu_3 & \mu_4 \\ \cline{3-4} \end{array}$$

Interaction is present when the effects of one independent variable depend on the levels of the other independent variable. Thus, an interaction is present when the difference between cell means μ_1 and μ_2, which represents the effect of factor A at one level of factor B (level b_1), *does not equal* the difference between cell means μ_3 and μ_4, which represents the effect of factor A at the other level of factor B (level b_2). That is,

$$\text{An interaction is present when } \mu_1 - \mu_2 \neq \mu_3 - \mu_4$$

(Again, as a reminder, the issue of "presence" or "absence" will be determined by an appropriate statistical test to account for differences that are due to chance.) Alternatively, we can define interaction in terms of the other independent variable: an interaction is present when the difference between cell means μ_1 and μ_3 (the effect of factor B at level a_1) does not equal the difference between cell means μ_2 and μ_4 (the effect of factor B at level a_2). That is,

$$\text{An interaction is present when } \mu_1 - \mu_3 \neq \mu_2 - \mu_4$$

Finally, another way of considering interaction is to look at the cell means on the *diagonals* of the 2 × 2 matrix. In this case, interaction is present when the sum of the cell means on one diagonal is not equal to the sum of the cell means on the other diagonal. That is,

$$\text{An interaction is present when } \mu_1 + \mu_4 \neq \mu_2 + \mu_3$$

All three ways of expressing interaction in a 2 × 2 matrix are equivalent.

Factorial designs and the assessment of interaction are valuable to any scientific enterprise, since they reveal how independent variables *combine* to influence behavior. More complex factorials in which three or more independent variables are manipulated yield extensive information on the interaction of independent variables. Ultimately, factorial designs can be used to provide a comprehensive picture of the behavior under study.

It is instructive to consider an example in which interaction is entirely absent. Consider the matrix of means in Table 13-3. First, you should note that we have chosen numbers for this example to duplicate exactly the two sets of marginal means from Table 13-2. What has changed in this example is the means within the matrix. Consider the pattern of differences revealed by the rows in Table 13-3; for example, between the means for physical science and social science, *exactly the same difference* is found for the marginal means (40.00 − 26.00 = 14.00 words) as for the computer presentation (47.83 − 33.83 = 14.00) and the standard presentation (32.17 − 18.17 = 14.00). Similarly, the difference between the marginal means for physical science and history (40.00 − 34.50 = 5.50) is identical to the difference found between means for the computer presentation (47.83 − 42.33) and for the standard presentation (32.17 − 26.67). We find the same outcome if we examine the other independent variable. That is, the difference between the marginal means for computer and standard presentations is 41.33 − 25.67 = 15.66, and exactly the same difference is found for all three lectures (47.83 − 32.17, 33.83 − 18.17, and 42.33 − 26.67).

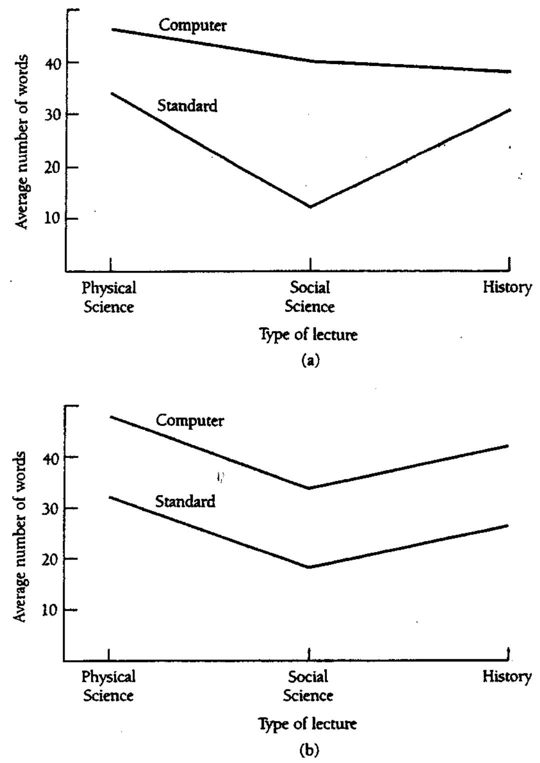
Table 13-3
An Example of No Interaction

Method of Presentation (Factor B)	Type of Lecture (Factor A)			Average
	Physical Science (a_1)	Social Science (a_2)	History (a_3)	
Computer (b_1)	47.83	33.83	42.33	41.33
Standard (b_2)	32.17	18.17	26.67	25.67
Average	40.00	26.00	34.50	

It is obvious in this case that the effects of the two independent variables, as reflected by the actual treatment means, are perfectly reflected in the marginal means, and that there is no interaction. In the case where interaction is absent, then, the influence of either variable is *not* dependent on the levels of the other.

The presence or absence of interaction is also effectively revealed by the kind of pictorial representation of the outcome of an experiment found in Fig. 13-1. The six treatment means from the original example (Table 13-2) are plotted in Fig. 13-1(a), while those from the second example (Table 13-3) are plotted in Fig. 13-1(b).[1] In each part of Fig. 13-1, the means for the levels of factor B are connected by separate lines. In such a graphic representation, interaction will be revealed by the presence of *nonparallel* lines, as in Fig. 13-1(a), while the absence of interaction results in *parallel* lines, as in Fig. 13-1(b). We will now turn to the statistical assessment of interaction (and of main effects), first by ANOVA and then by MRC.

[1] For convenience, we have treated factor A as a continuous independent variable, which of course it is not. This method of plotting data is commonly used by researchers, however, as it reveals the presence (or absence) of interaction quite clearly.



(a)

(b)

Figure 13-1 Two possible outcomes of the factorial experiment. (a) Graph displaying an interaction between the two independent variables. (b) Graph indicating that there is no interaction.

13.2 THE FACTORIAL ANALYSIS: THE ANOVA APPROACH

The analysis of variance is based on a partitioning of the total sum of squares into a number of component sums of squares, each of which reflects useful sources of variation. We will begin by describing these sources; as in the single-factor design, they may be expressed quite simply as deviations from means.

Sources of Variability

For the single-factor design, the deviation of each observation Y from the grand mean $\bar{Y}_T$ was divided into two portions: the deviation of the observation from the relevant treatment mean $\bar{Y}_A$ and the deviation of that treatment mean from the grand mean. In symbols,

$$Y - \bar{Y}_T = (Y - \bar{Y}_A) + (\bar{Y}_A - \bar{Y}_T)$$

You will recall that the first deviation to the right of the equal sign formed the basis for the within-groups sum of squares, and the second deviation, the basis for the between-groups sum of squares.

We use analogous partitioning of the deviations with the two-factor design. The total deviation may again be divided into the deviation of each observation Y from the relevant treatment mean—in the two-factor design, $\bar{Y}_{AB}$, which represents the mean for a combination of levels of the two independent variables—and the deviation of this treatment mean from the grand mean. In symbols,

$$Y - \bar{Y}_T = (Y - \bar{Y}_{AB}) + (\bar{Y}_{AB} - \bar{Y}_T)$$

As in the single-factor design, these two components form the basis for the within-groups and between-groups sums of squares, respectively. The within-groups sum of squares, which continues to reflect unsystematic variability—the uncontrolled variability of subjects treated alike—will be used to form the error term in the analysis of variance. The between-groups sum of squares, on the other hand, reflects several sources of systematic variability, which we will now isolate by further partitioning.

In Sec. 13.1 we indicated that these sources of variability are the main or average effects of the two independent variables and the effects of interaction between the variables. For the two main effects, the deviations involve the relevant column and row marginal means, namely,

$$\bar{Y}_A - \bar{Y}_T \quad \text{and} \quad \bar{Y}_B - \bar{Y}_T$$

for A and B, respectively. The deviation representing interaction is derived from the deviations we have already specified. Interaction may be viewed as the variability between groups that is *not* attributed to either of the two main effects. Thus, the interaction deviation is given by

$$(\bar{Y}_{AB} - \bar{Y}_T) - (\bar{Y}_A - \bar{Y}_T) - (\bar{Y}_B - \bar{Y}_T)$$

which simplifies to

$$\bar{Y}_{AB} - \bar{Y}_A - \bar{Y}_B + \bar{Y}_T$$

when we remove the parentheses and cancel out two of the grand means. To summarize, the deviation of each observation from the grand mean may be divided into four components: a within-groups deviation, and deviations attributed to the main effects of A and B and to the $A \times B$ interaction; in symbols,

$$Y - \bar{Y}_T = (Y - \bar{Y}_{AB}) + (\bar{Y}_A - \bar{Y}_T) + (\bar{Y}_B - \bar{Y}_T) + (\bar{Y}_{AB} - \bar{Y}_A - \bar{Y}_B + \bar{Y}_T) \quad (13\text{-}1)$$

If the partitioning specified in Eq. (13-1) is accomplished for each observation and the deviations are squared and summed, the relationship between the components may be expressed in terms of sums of squares:

$$SS_T = SS_{S/AB} + SS_A + SS_B + SS_{A \times B} \quad (13\text{-}2)$$

Note that the within-groups sum of squares is designated by the subscript S/AB, which clearly specifies the nature of this sum of squares—the variability of subjects within their treatment groups.[2] Our next step is to examine the computational formulas for calculating these sums of squares.

Computational Formulas

You will recall from Chap. 6 that sums of squares based on particular deviations are easily formed by combining certain basic ratios in patterns that reflect the components of those deviations (see Sec. 6.2, under "Basic Ratios"). We will use Eq. (13-1) to perform a similar function for the present design.

Notation. Lowercase letters are used to designate certain numbers relevant to a specific experiment:

a = the number of levels of factor A.
b = the number of levels of factor B.
s = the sample size (the number of subjects randomly assigned to each of the different combinations of the levels of the two independent variables).

We will refer to specific levels of independent variables and combinations of levels with lowercase letters and numerical subscripts. Levels of factor A will be designated a_1, a_2, etc., while levels of factor B will be designated b_1, b_2, etc. Specific treatment combinations are designated by the appropriate levels of the two variables. For example, a_1b_2 refers to the treatment group receiving level a_1 in conjunction with level b_2, and a_2b_1 to the pairing of levels a_2 and b_1.

[2] As in the single-factor design, the $SS_{S/AB}$ consists of the variability of subjects treated alike, pooled over all of the treatment groups. There are six treatment groups in the present example, which means that $SS_{S/AB}$ is the sum of the six within-group sums of squares.

All the necessary calculations involve various sums and subtotals, which we will represent with capital letters:

Y = the individual observation or score.
AB = the subtotal for any one of the treatment groups; when needed, subscripts are used to specify the particular levels of the two factors a given group has received. For example, $(AB)_{1,2}$ = the sum of the scores for subjects receiving the combination of levels a_1 and b_2, and $(AB)_{2,1}$, the sum associated with the combination of levels a_2 and b_1.
A = the sum of all the AB sums for a particular level of factor A. Subscripts, again, may be used to designate specific levels.
B = the analogous sum for factor B.
T = the grand sum of the scores.

The Preliminary Analysis. The first step in the analysis usually consists of summing the Y scores and their squared values for each of the $(a)(b)$ treatment groups or combinations. The two resulting sets of sums may be used to calculate the usual descriptive statistics, the group means and the standard deviations. In addition, the two sets of sums are used to calculate the basic ratios entering into the calculation of the factorial sums of squares. The first set (the AB sums) are entered into a special matrix used to facilitate the calculation of basic ratios, while the second set (the sum of the squared Y scores) are simply combined to form one of the basic ratios required to calculate $SS_{S/AB}$ and SS_T.

The *AB* Matrix. We continue the analysis by entering the AB treatment sums according to the levels of the two independent variables into what we call an ***AB* matrix.** An AB matrix based on the sums from Table 13-2 is presented in Table 13-4. The column marginal totals (A) are formed by adding the AB sums within the individual columns of the matrix, and the row marginal totals (B) are similarly

Table 13-4
AB Matrix of Sums

	a_1	a_2	a_3	Sum
b_1	276	240	228	744
b_2	204	72	186	462
Sum	480	312	414	1206

Table 13-5
Basic Ratios and Analysis Summary

Basic Ratios

$$[Y] = \Sigma\, Y^2 = 53^2 + 49^2 + \cdots + 30^2 + 20^2 = 46{,}194$$

$$[AB] = \frac{\Sigma\,(AB)^2}{s} = \frac{276^2 + 204^2 + \cdots + 228^2 + 186^2}{6} = \frac{267{,}156}{6} = 44{,}526.00$$

$$[A] = \frac{\Sigma\, A^2}{(b)(s)} = \frac{480^2 + 312^2 + 414^2}{(2)(6)} = \frac{499{,}140}{12} = 41{,}595.00$$

$$[B] = \frac{\Sigma\, B^2}{(a)(s)} = \frac{744^2 + 462^2}{(3)(6)} = \frac{766{,}980}{18} = 42{,}610.00$$

$$[T] = \frac{T^2}{(a)(b)(s)} = \frac{1206^2}{(3)(2)(6)} = \frac{1{,}454{,}436}{36} = 40{,}401.00$$

Summary of the Analysis

Source	*SS*	*df*	*MS*	*F*
A	[A] − [T] = 1194.00	2	597.00	10.74*
B	[B] − [T] = 2209.00	1	2209.00	39.73*
A × *B*	[AB] − [A] − [B] + [T] = 722.00	2	361.00	6.49*
S/AB	[Y] − [AB] = 1668.00	30	55.60	
Total	[Y] − [T] = 5793.00	35		

*$p < .01$.

formed within the rows. The grand total T is obtained by summing either set of marginal totals. We are now ready to calculate the basic ratios.

Basic Ratios. Formulas for the five required basic ratios, of which four are based on the sums appearing in the AB matrix and the fifth on the individual Y scores, are presented in the upper portion of Table 13-5. The first basic ratio we will consider is simply the sum of all of the squared scores. The formula for this ratio and the expansion based on the data from Table 13-2 are presented in the first row of the table.

The other basic ratios are based on the different sets of sums appearing in the AB matrix. In all cases, the numerator is formed by squaring the entire set of relevant sums (either AB, A, B, or T) and then summing the squares. In symbols, the numerators are:

$$\Sigma\,(AB)^2, \quad \Sigma\, A^2, \quad \Sigma\, B^2, \quad T^2$$

Each denominator consists of the number of observations contributing to the squared terms in the corresponding numerator:

s for the basic ratio based on the AB sums
$(b)(s)$ for the basic ratio based on the A sums
$(a)(s)$ for the basic ratio based on the B sums
$(a)(b)(s)$ for the basic ratio based on T

(The denominator for the basic ratio, which is based on the individual scores, Y, is 1 and does not need to be specified.)

The completed ratios and relevant calculations performed on the data from the numerical example are found in the remaining rows of the upper portion of Table 13-5. For convenience, each ratio is uniquely coded in order to simplify computational formulas for the different sums of squares.

Sums of Squares. The computational formulas for the sums of squares, which are presented in the bottom portion of Table 13-5, are specified in terms of the basic ratios. You will note that the patterns of combination are identical to the patterns of the components of the deviations upon which the sums of squares are based; see Eq. (13-1). The results of the operations are given in the column labeled *SS*. As an arithmetic check, you should verify that the sum of the component sums of squares equals the total sum of squares:

$$\begin{aligned} SS_T &= SS_A + SS_B + SS_{A \times B} + SS_{S/AB} \\ &= 1194.00 + 2209.00 + 722.00 + 1668.00 \\ &= 5793.00 \end{aligned}$$

The Analysis of Variance

The final steps in the calculations consist of determining the degrees of freedom for each source, calculating the mean squares, and forming the three F ratios for the effects analyzed. These steps are summarized in the remaining columns of the table.

Degrees of Freedom. The degrees of freedom for any main effect are simply the number of levels for each factor less 1. In this case,

$$df_A = a - 1 = 3 - 1 = 2$$
$$df_B = b - 1 = 2 - 1 = 1$$

The degrees of freedom for the $A \times B$ interaction are found by multiplying the df's for the two main effects. That is,

$$df_{A \times B} = (df_A)(df_B) = (2)(1) = 2$$

The degrees of freedom for the within-groups source ($df_{S/AB}$) are found by pooling the df's for all the treatment groups. The degrees of freedom for any one group are $s - 1$. Since there are $(a)(b)$ groups,

$$\begin{aligned} df_{S/AB} &= (a)(b)(s-1) \\ &= (3)(2)(6-1) = (6)(5) = 30 \end{aligned}$$

Finally, the degrees of freedom for SS_T are 1 less than the total number of observations, $(a)(b)(s)$; in symbols

$$\begin{aligned} df_T &= (a)(b)(s) - 1 \\ &= (3)(2)(6) - 1 = 36 - 1 = 35 \end{aligned}$$

As a check, df_T should equal the sum of the component df's:

$$\begin{aligned} df_T &= df_A + df_B + df_{A \times B} + df_{S/AB} \\ &= 2 + 1 + 2 + 30 = 35 \end{aligned}$$

Mean Squares and *F* Ratios. Any mean square is calculated by dividing a sum of squares by its degrees of freedom. The mean squares for the analysis are presented in Table 13-5. The F ratios are found by dividing the mean squares representing the factorial effects of interest by the within-groups mean square:

$$F = \frac{MS_{effect}}{MS_{S/AB}}$$

The F ratios for the two main effects and for the interaction are listed in the last column of the table.

The statistical hypotheses underlying the F tests are pairs of null and alternative hypotheses. For the main effects, the null hypothesis states that the population treatment means corresponding to the separate main effects are the same; the alternative hypothesis states that they are not all equal. For the interaction, the null hypothesis states that interaction effects are completely absent in the population; the alternative hypothesis states that they are not.

The logic behind these F tests is the same as that described for the single-factor design. Each numerator provides a population estimate of one of the three factorial effects *plus* error variance, while the denominator provides an estimate of error variance alone. Under the null hypothesis, both numerator and denominator mean squares reflect error variance and the expected value of each of the three F ratios is approximately 1.0. A significant F indicates that the null hypothesis is untenable and that we should accept the alternative hypothesis that a particular factorial effect—A main effect, B main effect, or interaction—is present.

We evaluate each null hypothesis by comparing the value of F we calculate and the critical value found in Table A-1, which, as usual, is determined by the df's associated with the numerator and denominator terms and the significance level chosen for the statistical tests. At $\alpha = .05$, the critical value for the A main effect and the $A \times B$ interaction is $F(2,30) = 3.32$ and that for the B main effect is $F(1,30) = 4.17$. An inspection of values in Table 13-5 indicates that all three factorial effects are significant. This means that:

1. There are differences overall among the three types of lecture.
2. There are differences overall between the two methods of presentation.
3. The differences among the three types of lecture depend on the method of presentation.

13.3 THE FACTORIAL ANALYSIS: THE MRC APPROACH

The identical statistical outcomes found with ANOVA for the two-factor design can be obtained with MRC. The researcher's critical step is in establishing vectors that capture the variation in Y that is of specific interest. The sources of this variation are, of course, the same sources that we normally isolate and study in an analysis of variance. We will consider the coding process first and then show how the factorial effects may be evaluated with MRC.

Coding of Vectors

Contrast coding is easily adapted to the analysis of a factorial experiment. The strategy we will follow is to establish sets of vectors for each of the two main effects and then use these vectors to define the vectors for the interaction.[3]

Coding Main Effects. We code each main effect by disregarding the levels of the other independent variable (or main effect) and then treating the main effect of interest as if it represented an independent variable in a *single*-factor design. The main effect of factor A, for example, requires two vectors ($df_A = 2$) to capture this variation, since there are three levels and we need $a - 1$ vectors in such a

[3] The method we recommend is not the only way to extract information about the main effects and interaction. On the other hand, its advantage, as you will see, is the ease with which the factorial analysis can be accomplished and the fact that it focuses on the sorts of meaningful questions we will consider in subsequent chapters.

In Chap. 13, you saw how you could study the influence of two independent variables that were being manipulated simultaneously in the context of a single experimental design. The critical question surrounding the analysis of this joint manipulation is whether the two independent variables *interact* to influence behavior. If they do, you need to examine the influence of each independent variable with the specific levels of the other variable clearly in mind. On the other hand, if they do not interact, you may examine the influence of either variable without reference to the other independent variable. The analyses covered in Chap. 13—the evaluation of the two main effects and that of the interaction—assess overall or *omnibus* effects: they indicate only whether a main effect or an interaction is present, not what is responsible for the significant *F*.

You will recall that we faced a similar problem with the omnibus test in the analysis of the single-factor design. The solution then was to locate meaningful differences between means. We look for the same sort of information with the factorial design, except that the procedures are more complicated, and which procedures we use depends on whether significant interaction is present. If there is a significant interaction, for example, we tend to focus our attention on the *individual treatment means* within the body of the *AB* matrix, our goal being to establish how the effects of one independent variable *change* with the different levels of the other independent variable. Without a significant interaction, we turn instead to the *marginal means*—in effect, treating the design as two separate *single*-factor experiments.

In Sec. 14.1 we will consider significant main effects, not because they are more important, but because the analysis is easily generalized from the single-factor design. The analysis of interaction, which we consider in Secs. 14.2 through 14.4 and in Chap. 15, takes two forms. The first consists of a systematic examination of the data row by row or column by column in the *AB* matrix in an attempt to establish the nature of the interaction. This approach is called **analysis of the simple effects** of an interaction. It involves an examination of the effects of *one* of the independent variables while the other independent variable is held constant. For example, we would look at the differences in vocabulary scores for the different types of lectures, but only under one method of presentation at a time—the computer *or* the standard method. The second approach, which we will consider in Chap. 15, consists of an examination of smaller factorial designs constructed from the larger design in order to express interaction in terms of more focused manipulations. We will call this approach **analysis of interaction comparisons**. In essence, this analysis takes a multilevel factorial design (e.g., a 4 × 3 design) and reduces it ideally to a number of 2 × 2 designs. Both types of analyses are useful, and we need to master both in order to understand fully the wealth of information available from the results of a factorial experiment.

14.1 DETAILED ANALYSIS OF MAIN EFFECTS

If the interaction is not significant, the design becomes for all practical purposes *two* single-factor designs. That is, we examine each factor alone—looking at the differences among the ***row marginal means*** and the differences among the ***column marginal means*** separately—without reference to the levels of the other independent variable. Only main effects associated with more than 1 *df* are candidates for further analysis, of course, since a main effect with 1 *df* is already a difference between two marginal means. In our present example, then, the *B* main effect, consisting of two levels and based on 1 *df*, reflects the difference between computer presentation and the standard presentation; no further analysis is possible. The *A* main effect, consisting of three levels and based on 2 *df*, reflects the undifferentiated effects of the three different lectures; additional analyses are necessary to determine which comparisons between treatments are useful and interesting.

The ANOVA Approach

The analysis under ANOVA consists of a simple extension of the procedures described for the single-factor design. We first express the comparison ($\hat{\psi}_A$), which is usually a difference between two means, in terms of coefficients c_i. That is,

$$\hat{\psi}_A = \Sigma\ (c_i)\,(\bar{Y}_{A_i}) \qquad (14\text{-}1)$$

This comparison is then combined with other quantities to calculate the sum of squares associated with it:

$$SS_{A_{comp.}} = \frac{(b)\,(s)\,(\hat{\psi}_A)^2}{\Sigma\ (c_i)^2} \qquad (14\text{-}2)$$

This formula is identical to Eq. (11-4), the formula for the single-factor design, except for the insertion of *b* in the numerator to reflect the number of observations contributing to each mean. More specifically, there were *s* observations associated with each $\bar{Y}_A$ in the single-factor design, and there are $(b)(s)$ observations associated with each of the means contributing to the main effect of factor *A* in the two-factor design. The error term for these single-*df* comparisons is the within-groups error term from the overall analysis, $MS_{S/AB}$. We use this error term because it captures *Y* variability based on the full set of available data.

We will use the data from Chap. 13 for a numerical example. The marginal means are 40.00, 26.00, and 34.50 for the lectures on physical science, social science, and history, respectively. Suppose we wish to compare the average of the means for the two science lectures with the mean for the history lecture. For this comparison, we will use the coefficients $(+\frac{1}{2}, +\frac{1}{2}, -1)$. With these coefficients, we can calculate the difference between the combined science mean and the history mean:

$$\begin{aligned}\hat{\psi}_A &= (+\tfrac{1}{2})(40.00) + (+\tfrac{1}{2})(26.00) + (-1)(34.50)\\ &= 33.00 - 34.50 = -1.50\end{aligned}$$

This value is substituted in Eq. (14-2) to obtain

$$\begin{aligned}SS_{A_{comp.}} &= \frac{(2)(6)(-1.50)^2}{(+\frac{1}{2})^2 + (+\frac{1}{2})^2 + (-1)^2}\\ &= \frac{27.00}{1.50} = 18.00\end{aligned}$$

Since this sum of squares is based on 1 *df*, $MS_{A_{comp.}} = 18.00$. The *F* ratio is:

$$F = \frac{MS_{A_{comp.}}}{MS_{S/AB}}$$

From Table 13-5, we find that $MS_{S/AB} = 55.60$. Completing the operations, we find

$$F = \frac{18.00}{55.60} = .32$$

This difference is not significant. The *df*'s for this *F* are $df_{num.} = 1$ and $df_{denom.} = df_{S/AB} = 30$. (Be sure to note that the denominator *df* are those associated with the error term from the omnibus analysis, $df_{S/AB}$.)

Single-*df* comparisons involving the main effect of factor *B* (which are not possible in the present example because $df_B = 1$) would be calculated in the same manner. All one needs to do is to modify Eq. (14-2) to reflect the appropriate number of observations contributing to each marginal mean, namely, $(a)(s)$. Thus,

$$SS_{B_{comp.}} = \frac{(a)(s)(\hat{\psi}_B)^2}{\Sigma\,(c_j)^2} \qquad (14\text{-}3)$$

From this point on, we would follow exactly the same steps we outlined above for comparing the *A* treatment means.

The MRC Approach

The detailed analysis of main effects is easily accomplished with MRC, provided we use meaningful comparisons as vectors. For our example, consider the first two vectors we established for the data set in Chap. 13 (Table 13-6), which we used to capture the *A* main effect. The first vector (*A*1), based on the coefficients (+1, −1, 0), specifies a comparison between physical science and social science, while the second vector (*A*2), based on the coefficients (+1, +1, −2), specifies a comparison between the combined science conditions and the history condition.

This latter comparison is the same comparison we just considered in the first part of this section, "The ANOVA Approach." All that we have to do now is to locate the appropriate zero-order correlation between *Y* and *A*2 from the MRC analysis and test its significance.

From Table 13-6, we find $r_{Y2}^2 = .0031$. To test the significance of this r^2, we calculate the *F* ratio using Eq. (11-9):

$$F = \frac{r^2_{comp.}}{(1 - R^2_{Y.max.})/(N - k - 1)}$$

where $r^2_{comp.}$ is the squared zero-order correlation coefficient representing the single-*df* comparison and $R^2_{Y.max.}$ is the squared *omnibus* multiple correlation coefficient, obtained when the required full set of vectors (the vectors for the design—*A*, *B*, and *A* × *B*, or 5 vectors) is entered into the analysis. The numerator of the *F* ratio reflects the proportion of *Y* variability associated with a single vector, which in this case represents a single-*df* comparison. The denominator contains the quantity $1 - R^2_{Y.max.}$, which is the proportion of *Y* variability not associated with the combined effects of the experimental treatments. This residual term is divided by the appropriate *df*, which corresponds, of course, to the *df* for the within-groups source of variance, $df_{S/AB}$. (In MRC, as you know, these *df* are represented by $N - k - 1$, where *N* is the total number of observations and *k* is the maximum number of vectors required to capture the combined treatment effects.)

All that we need to complete this example is the residual term and the residual *df*, which may be found in Table 13-8—namely, $1 - R^2_{Y.max.} = .2880$ and $df_{S/AB} = N - k - 1 = 30$. If we substitute the different quantitites in Eq. (11-9), we obtain

$$F = \frac{.0031}{.2880/30} = \frac{.0031}{.0096} = .32$$

which is identical to the *F* we obtained with ANOVA for this comparison.

The correspondence between ANOVA and MRC, again, can be demonstrated in other ways as well. The $SS_{A_{comp.}}$ can be obtained from the MRC analysis simply by multiplying the $r^2_{comp.}$ by SS_T. Thus,

$$SS_{A_{comp.}} = (r^2_{comp.})(SS_T) = (.0031)(5793.00) = 17.96$$

which, except for rounding error, is equal to the value of 18.00 obtained with ANOVA. The $r^2_{comp.}$ may also be calculated from ANOVA by dividing the $SS_{A_{comp.}}$ by SS_T. That is,

$$r^2_{comp.} = \frac{SS_{A_{comp.}}}{SS_T} = \frac{18.00}{5793.00} = .0031$$

which is equal to the squared correlation coefficient obtained with MRC.

Comment

You have seen how differences among the marginal means may be analyzed by either statistical approach. With ANOVA, the focus is on the difference between the means represented by the single-*df* comparison of interest. With MRC, the focus is on the zero-order correlation between *Y* and the vector reflecting the single-*df* comparison. An obvious strategy in planning the MRC analysis is to represent the main effects of the two independent variables by meaningful comparisons. This way the relevant zero-order correlations are readily available for a detailed analysis of the main effects—if such an analysis is appropriate, of course. If you desire additional comparisons, you can calculate these $r_{comp.}$'s by simply including appropriate vectors when you set up the data for analysis and instructing the computer to include these vectors in the zero-order correlation matrix available from the MRC program.

As we pointed out earlier, systematic interest in differences among the marginal means generally is relevant only when the interaction is not significant. The primary reason has to do with *interpretation*: differences among marginal means are often difficult to interpret when there is a significant interaction—as are the main effects themselves. When interaction is present, any conclusion drawn from an analysis of the marginal means will need to be qualified. Consider the data from our numerical example presented in the upper half of Fig. 13-1. The analysis revealed a significant main effect of presentation method. Although it would probably be safe to conclude that the computer method was generally superior to the standard method, since the computer method was consistently better for all three lectures, we would still have to take into consideration the fact that the *size* of its superiority depends on the type of lecture presented—large for social science and small for physical science and history—whenever we interpret the results of the experiment.

14.2 USING ANOVA TO ANALYZE SIMPLE EFFECTS

General Considerations

Once a significant interaction has been established, researchers usually turn their attention to an analysis of the means *within* the body of the *AB* matrix; they have little interest in the analysis of the marginal means. One commonly used technique consists of the systematic analysis of the treatment means—either one row at a time or one column at a time.[1] This type of analysis is called the analysis of **simple main effects**, or **simple effects**, for short.

The analysis of simple effects consists of the examination of the effects of *one* of the independent variables with the *other* independent variable *held constant*. In the context of our example, we might examine separately the effects of the different lectures under computer presentation and then their effects under the standard presentation. In each situation, we are holding presentation method constant while permitting only the type of lecture to vary in the analysis.

In essence, then, this is an analysis scheme that views the factorial design as a collection of *separate single-factor experiments*, each involving the same manipulation. Because there is a significant interaction, we can conclude that the outcomes of these "separate experiments" are in fact *not the same*; the analysis of simple effects represents an attempt to determine the ways in which these outcomes differ. It is in this sense that we come to "understand" or "explain" a significant interaction through the analysis of simple effects.

The analysis described above focused on the simple effects of factor *A*, which consisted of the effects of types of lectures for the computer presentation (referred to as the *simple effects of A at level* b_1) and for the standard presentation (referred to as the *simple effects of A at level* b_2). We could just as well have considered the effects of the two methods of presentation (factor *B*) separately at each level of factor *A*. In this case, there would be three "single-factor experiments" all involving the comparison of computer and standard presentation, but with the type of lecture held constant. Specifically, there would be one such "experiment" for the subjects receiving the physical science lecture, one for the subjects receiving the social science lecture, and one for the subjects receiving the history lecture. These analyses would be called the *simple effects of factor B at levels* a_1, a_2, *and* a_3, respectively.

Analysis of Simple Effects

Since the specification of a simple effect is equivalent to a single-factor experiment, the analysis builds on procedures we have already considered in earlier chapters. We isolate the appropriate column or row in the *AB* matrix and calculate a sum of squares based only on the data in it. We then treat this subset of the data, which represents the comparison of interest, exactly as if it had come from an actual single-factor experiment, rather than from a "slice" or a part of the *AB* matrix.

As an example, the treatment sums for the three lectures (factor *A*) given with the computer presentation (level b_1) are:

[1] Occasionally, it is profitable to conduct the analyses both ways, i.e., by rows and by columns.

a_1	a_2	a_3	Sum
276	240	228	744

Each sum is based on $s = 6$ observations. If we treat these data as the results from a single-factor experiment, the between-groups sum of squares ("SS_A") is calculated by:

$$SS_{A \text{ at } b_1} = \text{“}SS_A\text{”}$$
$$= \frac{\Sigma \text{“}A\text{”}^2}{s} - \frac{\text{“}T\text{”}^2}{(a)(s)}$$

where "A" and "T" are data drawn from one of the rows of the AB matrix. Substituting in this formula, we find

$$SS_{A \text{ at } b_1} = \frac{276^2 + 240^2 + 228^2}{6} - \frac{744^2}{(3)(6)}$$
$$= \frac{185{,}760}{6} - \frac{553{,}536}{18}$$
$$= 30{,}960.00 - 30{,}752.00 = 208.00$$

The df associated with this sum of squares is 1 less than the number of treatment conditions; that is,

$$df_{A \text{ at } b_1} = a - 1 = 3 - 1 = 2$$

The mean square is formed in the usual manner by dividing the SS by the appropriate df:

$$MS_{A \text{ at } b_1} = \frac{SS_{A \text{ at } b_1}}{df_{A \text{ at } b_1}}$$
$$= \frac{208.00}{2} = 104.00$$

The F is calculated by dividing the mean square representing systematic variance by the *error term from the overall omnibus analysis* (see Table 13-5). For these data,

$$F = \frac{MS_{A \text{ at } b_1}}{MS_{S/AB}}$$
$$= \frac{104.00}{55.60} = 1.87$$

which is not significant. (With $df_{num.} = 2$ and $df_{denom.} = 30$, the critical value of F at $\alpha = .05$ is 3.32.) It appears that the three types of lectures produce roughly equivalent results under the computer presentation.

The corresponding analysis for the standard presentation produces a different conclusion. The treatment sums at level b_2 are:

a_1	a_2	a_3	Sum
204	72	186	462

Following the same steps, we find

$$SS_{A \text{ at } b_2} = \frac{204^2 + 72^2 + 186^2}{6} - \frac{462^2}{(3)(6)}$$
$$= \frac{81{,}396}{6} - \frac{213{,}444}{18}$$
$$= 13{,}566.00 - 11{,}858.00 = 1708.00$$
$$MS_{A \text{ at } b_2} = \frac{1708.00}{2} = 854.00$$
$$F = \frac{854.00}{55.60} = 15.36$$

where the value of F is significant.

Thus, the analysis shows that the interaction may be characterized as consisting of a nonsignificant effect of lectures under computer presentation and a significant effect of lectures under the standard presentation. As informative as this conclusion may be, we still do not know exactly what differences are responsible for the *significant* simple effect. All that the analysis establishes is that differences exist among the three lectures, not *where* they exist. Additional analysis will be necessary to reveal this important information.

Simple Effects of Factor *B*

The analysis of the simple effects of factor B is conducted in an analogous fashion. We use the same two-step process of determining which simple effects are significant and—when we locate those that are—of testing specific differences between means. If the simple effects are each associated with 1 *df*, as they are in the numerical example, only the first step is possible. That is, the analysis will consist of comparing the two methods of presentation separately for each of the three lecture conditions, and no further analysis is possible.

Analysis of Simple Comparisons

With the single-factor design, the determination of a significant omnibus *F* is usually followed by a number of additional statistical tests designed to establish the critical factors responsible for the significant *F*. We follow exactly the same logic when we discover a significant simple effect in a factorial experiment. In our example, an examination of the treatment means for the standard presentation suggests that there is a sizable difference in performance between subjects receiving the social science lecture and those receiving the other two lectures. The social science subjects recalled 12.00 vocabulary words, while the physical science subjects recalled 34.00 words and the history subjects recalled 31.00 words. We can test this observation by considering two comparisons: one, between physical science and history, to establish the equivalence of these two conditions; and another, between social science (12.00) and the average of physical science and history (32.50), to establish the discrepancy between social science and the other two groups.

It is important to note at this point that to be of any analytical benefit, single-*df* comparisons should represent meaningful questions. The analysis suggested in the preceding paragraph does not give us much insight into the underlying reasons for the effect. What is the basis for comparing physical science and history? A more revealing comparison is between the two science conditions. We made this point previously when we first introduced single-*df* comparisons in the analysis of the single-factor design. The point is equally valid when we are analyzing significant simple effects.

The analysis of simple comparisons is merely an extension of the analysis of single-*df* comparisons that we applied to the single-factor design. We use coefficients to express the simple comparisons in which we are interested and calculate the sums of squares. As an example, suppose we wanted to compare the two science means. The coefficients for this comparison are (c_i: $+1, -1, 0$), and the corresponding means are 34.00, 12.00, and 31.00 for physical science, social science, and history, respectively. The difference between the two means ($\hat{\psi}$) is

$$\begin{aligned}\hat{\psi}_{A \text{ at } b_2} &= (c_1)(\bar{Y}_{A_1B_2}) + (c_2)(\bar{Y}_{A_2B_2}) + (c_3)(\bar{Y}_{A_3B_2})\\ &= (+1)(34.00) + (-1)(12.00) + (0)(31.00)\\ &= 34.00 - 12.00 = 22.00\end{aligned}$$

This difference is substituted in an equivalent of Eq. (11-4) as follows:

$$\begin{aligned}SS_{A_{comp.} \text{ at } b_2} &= \frac{(s)(\hat{\psi}_{A \text{ at } b_2})^2}{\Sigma (c_i)^2}\\ &= \frac{(6)(22.00)^2}{(+1)^2 + (-1)^2 + (0)^2}\\ &= \frac{2904.00}{2} = 1452.00\end{aligned}$$

Since there is 1 *df* associated with this comparison, $MS_{A_{comp.} \text{ at } b_2} = 1452.00$. The *F* ratio is formed by dividing the comparison mean square by the error term from the overall analysis ($MS_{S/AB}$) That is,

$$\begin{aligned}F &= \frac{MS_{A_{comp.} \text{ at } b_2}}{MS_{S/AB}}\\ &= \frac{1452.00}{55.60} = 26.12\end{aligned}$$

which, with $df_{num.} = 1$ and $df_{denom.} = 30$, is significant.

14.3 USING MRC TO ANALYZE SIMPLE EFFECTS

You will recall that with MRC the number of vectors required to represent any given source of variability fully is equal to the *df* associated with that source. For the simple effects of factor *A*, then, $a - 1$ vectors will be needed for each level of factor *B*; for the present example, the number of vectors is 2. For simple effects of factor *B*, $b - 1$ vectors will be needed for each level of factor *A*; for the present example, the number is 1.

Coding Simple Effects

The method we will use to construct vectors that reflect simple effects is based on the vectors we use to code the relevant *main effect*. This method follows a relatively simple procedure:

1. Construct vectors to define the relevant main effect.
2. Use a coefficient of 0 for all observations *not relevant* to a particular simple effect.

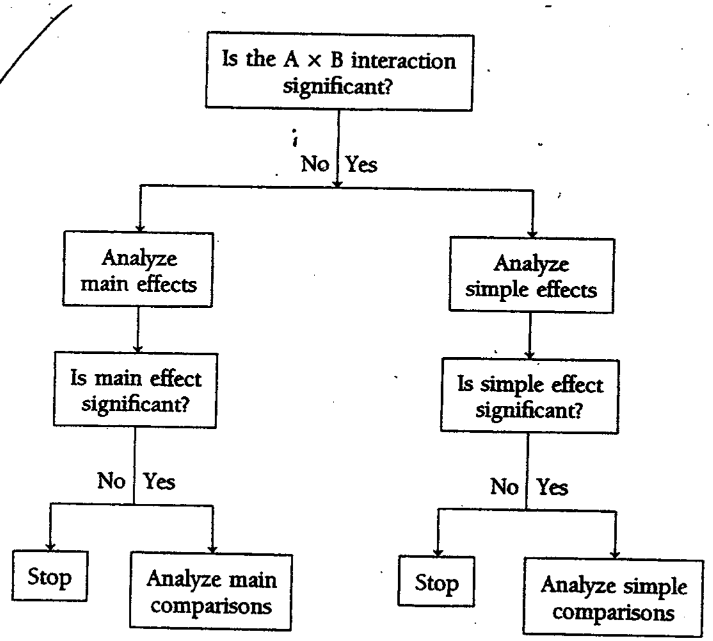

Is the A × B interaction significant?
No
Yes
Analyze main effects
Analyze simple effects
Is main effect significant?
Is simple effect significant?
No
Yes
No
Yes
Stop
Analyze main comparisons
Stop
Analyze simple comparisons

Figure 14-1 Schematic representation of an analysis with no planned comparisons.

All the experimental designs we have considered so far are examples of a particular class of designs in which subjects serve in only *one* of the treatment conditions. Because the assignment of subjects to conditions is random in these designs, they are called **completely randomized designs.** Since all treatment effects are based on differences between independent groups of subjects, these designs are also called **between-subjects designs.**

In another class of designs, one that is quite popular in the behavioral sciences, subjects serve in *several* or even *all* of the treatment conditions. Because different treatment effects observed in the same subjects represent differences *within* rather than *between* subjects, these designs are called **within-subjects designs.** (Such designs are also referred to as designs with **repeated measures.**) In this chapter, we will discuss the simplest within-subjects design, in which all subjects receive all levels of a single independent variable. In Chap. 17, we consider a relatively common factorial design in which only the levels of one of the independent variables are administered to the same group of subjects, while the levels of the other independent variable are administered to different groups of subjects.

16.1 ADVANTAGES AND DISADVANTAGES OF WITHIN-SUBJECTS DESIGNS

Advantages

There are several reasons why researchers choose within-subjects designs. One of these derives from the fact that these designs permit the examination of the effects of all levels of an independent variable at the level of the *individual subject.* Since each subject receives all the treatment conditions, we can study how each level of the independent variable affected each of the participants. All else being equal, researchers usually prefer to observe directly the effects of each level of the independent variable on individual subjects rather than inferring the effects of the levels from differences between groups of subjects receiving different treatments.

The within-subjects design is also ideally suited for studying such phenomena as learning, transfer of training, and practice effects of various sorts. In a learning experiment, for example, subjects are usually given repeated exposures on a particular task, and their performance is assessed following each practice trial. The researcher can then determine how subjects improve over repeated presentations of the task. The independent variable in this case consists of the number of learning trials given all subjects.

The primary reason why researchers choose within-subjects designs, however, is that such designs may help them increase the statistical *sensitivity*, or *power*, of the experiment. Under most circumstances, the error terms used to evaluate the significance of treatment effects in within-subjects designs are considerably *smaller* than those used in corresponding between-subjects designs. With smaller error terms, more treatment effects will be significant, and so the power will be increased. Why are error terms smaller in this type of design? We will consider a simple explanation before turning to the detailed analysis in later sections of this chapter.

Suppose we are comparing three treatment conditions in a between-subjects design. By now you realize that even sizable differences among the means might be entirely due to uncontrolled factors. A major source of uncontrolled variability is the fact that subjects with widely different abilities are randomly assigned to the treatment groups. As you have seen, the pooled within-groups mean square (for example, $MS_{S/A}$, for a single-factor experiment) provides an estimate of the degree to which the differences among the means may be reasonably viewed as only the result of uncontrolled subject differences.

Consider another comparison of three treatment conditions, this time set up as a within-subjects design in which subjects receive all three treatments. How are we now to interpret differences found among the treatment means? Can uncontrolled factors still affect the outcome of the experiment, or are they all eliminated because the same subjects are tested in all the treatment conditions? A moment's reflection should suggest that it is virtually impossible to remove completely the influence of uncontrolled factors in any experiment. There are still differences introduced by our inability to exactly duplicate the treatment conditions for different subjects; things that might vary from test to test include factors inherently associated with the treatments themselves, such as the calibration of any equipment and the exact reading of instructions; and other, external factors, such as room temperature, lighting, and background noises. Moreover, even the same subject will change slightly on repeated testing: variations in motivation, attitude, and other factors can cause inconsistent behavior in subjects. On the other hand, it will usually be true that the collective influence of all of these uncontrolled factors will be less with this design than with randomly formed groups of subjects. The result, then, is a smaller error term with which to evaluate the observed differences among treatment means. Assuming that the treatment effects are the same in the two designs, the within-subjects design will be more sensitive than the corresponding between-subjects design. This is because the *denominator* in the F ratio is smaller and, thus, the F ratio itself is larger.

How can the between-subjects design "compete" with this more sensitive design? First, there are circumstances in which testing subjects more than once is either not possible or not feasible. Any experiment using differential instructional sets to define the different treatments, for example, can probably not be convincingly administered to the same subjects. As another example, there are some experiments in which previously administered treatments continue to influence subjects' behavior under a new treatment condition. Second, there are experiments in which subjects are not willing or able to serve in more than one treatment condition. Human subjects who are serving in an experiment as part of a course requirement, for example, may have only an hour or two to spend; a complete administration of the treatments, on the other hand, may take considerably longer. Third, it is always possible to achieve an increase in power by adding to the sample size, rather than by selecting a within-subjects design. Of course, limits on time, money, and other resources may reduce the effectiveness of this strategy. Finally, between-subjects designs can often be made considerably more sensitive through the use of a statistical procedure called the *analysis of covariance.* (We consider this analysis in Chap. 22.)

Disadvantages

There are several problems associated with repeated measures, namely, practice effects, differential carryover effects, and the potential for violations of certain statistical assumptions. We will consider each briefly.

General Practice Effects. There is no foolproof way of escaping the fact that the performance of the subjects will change systematically during the course of receiving some or all of the treatments in an experiment. The changes may be either positive or negative. On the positive side, in experiments where familiarity with the experimental procedures increases performance, subjects will usually improve as they gain experience with the general requirements of the experiment. On the negative side, subjects may "deteriorate" on subsequent tests as they become bored or tired during the course of the experiment. We will refer to any such changes that occur during the course of testing as **practice effects.** As conceptualized, practice effects are assumed to be *general* and *not the result of exposure to any particular treatment condition* (or conditions). Since it is unlikely that these positive and negative effects of repeated testing will be in perfect balance, we must take their net effect into consideration when designing an experiment.

One solution is to introduce procedures that are designed to eliminate the general effects of repeated testing. If improvement with repeated trials is a possibility, for example, subjects can be given preliminary training on a relevant task before they receive the independent variable so that the improvement is unlikely to occur during the actual experiment. (This technique is frequently used in psychophysical studies and in experiments with animals.) As for the negative factors, boredom can often be minimized through the use of monetary incentives designed to maintain the same level of motivation during the course of the study, while fatigue can be reduced by introducing rest periods between successive administrations of the treatments.

Rarely will these steps remove all practice effects *completely*, however. For this reason, then, researchers usually employ an additional technique which spreads any remaining practice effects equally over the treatment conditions. To see how this works, consider a within-subjects design consisting of only two treatment conditions. Suppose there is a positive practice effect in this experiment, with subjects showing higher scores on whatever task they receive second. If the treatments are presented in the same order to all subjects—for example, condition 1 and then condition 2—it will not be possible to disentangle any effects produced by the different treatments from the overall improvement due to the practice effects. This is because performance on only one of the conditions (condition 2) will show the benefits from practice. We can avoid this problem quite simply by reversing the order of the two conditions (administering condition 2 and then condition 1) for half of the subjects. This way, scores in both conditions benefit equally from any practice effect, removing it as a potential source of bias.

A common technique for neutralizing the effects of practice is **counterbalancing,** which consists of the systematic variation of the order in which the treatments are presented to different subjects. This procedure, which is discussed in most elementary textbooks on experimental design, guarantees that each treatment condition is presented an equal number of times first, second, third, and so on, in particular sequences of conditions given different subjects.

Differential Carryover Effects. **Differential carryover effects** are lingering effects of one or more earlier treatment conditions that combine with the effects of treatments administered later in the testing order. Since these effects will rarely be the same for all conditions—which is why the word *differential* is used—they cannot be neutralized with counterbalancing. For this reason, therefore, they pose a serious problem for any researcher contemplating a within-subjects design.

Consider, for example, an experiment consisting of a drug condition and a control condition. We can reasonably expect that if subjects experience the drug condition first, its effect will influence how they behave when they subsequently receive the control condition, but that experiencing the control condition first will have little effect on how they respond to the subsequent drug condition. The only circumstance under which counterbalancing will work is when the carryover effects for all conditions and orders are the same. Consequently, counterbalancing will not eliminate carryover effects in this situation.

Problems of this sort can occur whenever the treatment conditions differ dramatically, as do the control condition and the experimental condition we discussed in the preceding paragraph. Similar problems often result when instructions are used to create a new treatment condition and to change a subject's perception of a task performed in a prior treatment condition—a common technique in behavioral research. It may be virtually impossible with instructional independent variables to have subjects completely disregard what they have been told in a previous condition. Greenwald (1976) provides a useful discussion of these sorts of problems, which he refers to as *context effects*.

It is often difficult to distinguish between practice effects and differential carryover effects, since both effects result from subjects' receiving more than one of the treatment conditions. The difference lies in whether changes with successive testing are the same for all conditions. In the drug example, they are not. You can always check for the presence of differential carryover effects by plotting the means for each treatment condition on a graph as a function of when it was administered—overall means for first, second, and third place, and so on—and then comparing the "practice" curves for the different conditions. Practice effects will be revealed if the functions for the different treatments exhibit the *same overall shape*; differential carryover effects will be revealed if the shapes are *different*.

You should always consider carefully the possibility of differential carryover effects whenever you contemplate a within-subjects design. Even if you do not expect differential carryover effects to appear, you should nevertheless always examine the treatment means for them. If they do appear, you can still compare the treatment conditions on the *first test*, of course, since carryover effects cannot appear until after the first test; but any statistical tests on the data will probably lack power because of the small number of subjects assigned to the different conditions.

Statistical Assumptions. The use of the F test to evaluate the significance of treatment effects is predicated on a number of assumptions. In addition to the assumptions of normality, homogeneity of within-treatment variances, and independence, which underlie the statistical analysis of completely randomized designs (see Sec. 8.4), within-subjects designs operate under an assumption concerning the correlations between the multiple measures obtained from the same subjects: that the correlations between all possible pairs of treatments are equal. With three treatments, the assumption is that the correlations between levels a_1 and a_2, between levels a_1 and a_3, and between levels a_2 and a_3 are equal. For the evaluation of F ratios with completely randomized designs (see Sec. 8.4) only severe violations concerning the nature of the distributions of treatment populations are critical, but such is not the case with within-subjects designs. Even minor violations of the underlying assumption will affect how we evaluate the significance of an F ratio in that the critical values of F obtained from Table A-1 are *too small*—the actual critical values we should be using are larger than those listed in the F table. A relatively simple solution to this problem is to use a slightly more stringent significance level—.025 rather than the standard .05—which will correct the difficulty in most situations.[1] Fortunately, evaluation of single-*df* comparisons seems to be unaffected by these violations, provided that specific error terms are used for these tests (Keppel, 1982, pp. 472–473). We will discuss the use of such error terms in Sec. 16.4.

16.2 THE OVERALL ANALYSIS: THE ANOVA APPROACH

For the overall statistical analysis of the single-factor within-subjects design, we make use of procedures we considered in earlier chapters. Only one point is new: in order to evaluate the significance of the treatment effects, we need to calculate an error term that takes into account that all subjects receive all the treatment conditions.

Design and Notation

The single-factor within-subjects design is defined as an experiment in which all subjects are tested under all the treatment conditions. In fact, the design can be viewed as a type of factorial design in which the independent variable (factor A) and *subjects* (factor S) are crossed to form all possible combinations of the levels of the two factors. We will refer to this arrangement as an $(A \times S)$ design to emphasize the relationship of this design to an actual factorial. (We use parentheses to designate a within-subjects factor, for reasons that we will explain in Sec. 17.1.)

As an illustration of the $(A \times S)$ design, consider an experiment in which $s = 3$ subjects are each tested under all $a = 3$ treatment conditions. We will assume that some form of counterbalancing is used in order to balance possible practice effects.[2] This design and the notation required are presented on the right-hand side of Table 16-1. For contrast, the correspondiing between-subjects design is presented on the left. It is important to note the differences between these two designs. Both designs produce the same quantity of data, namely, three Y scores obtained from each of the conditions. The basic difference is that the *same* three subjects are represented under each of the three treatments in the within-subjects design, while three *different* subjects are correspondingly represented in the between-subjects design. Because of this difference, we have one additional piece of information from the $(A \times S)$ design that we lacked with the other design, namely,

[1] The nature of these assumptions and the steps that can be taken to reduce the effects of violating them are discussed in most advanced statistical books (see, for example, Keppel, 1982, pp. 467–473; Kirk, 1982, pp. 256–262; Myers, 1979, pp. 171–174; and Winer, 1971, pp. 281–283).

[2] This could be accomplished, for example, by presenting the three conditions in the order 1-2-3 for the first subject, 2-3-1 for the second subject, and 3-1-2 for the third subject.

Table 16-1
Comparison of the Between-Subjects and Within-Subjects Designs

Between-Subjects Design						Within-Subjects Design				
a_1		a_2		a_3			a_1	a_2	a_3	Sum
s_1	$Y_{1,1}$	s_4	$Y_{2,4}$	s_7	$Y_{3,7}$	s_1	$Y_{1,1}$	$Y_{2,1}$	$Y_{3,1}$	S_1
s_2	$Y_{1,2}$	s_5	$Y_{2,5}$	s_8	$Y_{3,8}$	s_2	$Y_{1,2}$	$Y_{2,2}$	$Y_{3,2}$	S_2
s_3	$Y_{1,3}$	s_6	$Y_{2,6}$	s_9	$Y_{3,9}$	s_3	$Y_{1,3}$	$Y_{2,3}$	$Y_{3,3}$	S_3
Sum	A_1	Sum	A_2	Sum	A_3	Sum	A_1	A_2	A_3	T

an overall sum of the treatment scores for each subject. The sums for individual subjects, which are designated S_1, S_2, and S_3, are represented as row marginal sums in the table. As you will see, we will need these sums when we calculate the new error term.

The Analysis

With the between-subjects design, we divided the total sum of squares SS_T into two component parts, the treatment sum of squares SS_A and the within-groups sum of squares $SS_{S/A}$. This latter quantity, which is based on the pooled variation of subjects treated alike, is used to calculate the error term; it is an estimate of the uncontrolled variability present in an experiment based on a between-subjects design. As we have pointed out already, the $(A \times S)$ design provides a way of reducing this uncontrolled variability; it effectively reduces the contribution of chance factors to the differences among the treatment means. Using each subject for all the treatments allows us to obtain an estimate of the degree to which individual subjects respond consistently across the conditions, or an estimate of the consistency of individual differences. If we can assess this consistency, we have, in effect, explained more of the dependent variable, and thus leave a lower amount of unexplained variability.

Calculating the New Sums of Squares. The key to the analysis, then, is to estimate the degree to which using the same subjects represents a consistent or constant factor in the experiment. Such an estimate is easily calculated from the information provided by the overall sums for the respective subjects. That is, the sum for each subject (S_j) can be transformed into a mean for each subject ($\bar{Y}_{S_j}$), which in turn can be represented as a deviation from the overall mean $\bar{Y}_T$ and, ultimately, as a sum of squares SS_S. Stated another way, we can compute a main effect of subjects (SS_S), which represents the degree to which the subjects behave *consistently* as

they shift from treatment to treatment. If we subtract this sum of squares from the pooled "within-groups" sum of squares $SS_{S/A}$, which, you will recall, would represent uncontrolled variation if we ignored the fact that the same subjects are involved, we obtain a new sum of squares that can be used to estimate the degree to which uncontrolled factors are operating in the $(A \times S)$ design. That is,

$$SS_{error\ term} = SS_{S/A} - SS_S$$

We noted in the preceding paragraph that the subject sum of squares is based on the deviation of each subject's average score $\bar{Y}_S$ from the grand mean $\bar{Y}_T$; in symbols,

$$\bar{Y}_S - \bar{Y}_T$$

As you have seen before, we can use this deviation to express the sum of squares in terms of basic ratios. More specifically,

$$SS_S = [S] - [T]$$

where [S] and [T] represent basic ratios based on subject sums S and the grand sum T, respectively. The only new quantity is [S], which is calculated as follows:

$$[S] = \frac{\Sigma S^2}{a}$$

where S = the sum obtained by adding together all the Y scores for each subject

a = the number of treatments given to each subject

The formula for [T] should be familiar to you by now.

As we indicated earlier, the sum of squares for the error term used in the analysis of the $(A \times S)$ design may be obtained by subtraction. Expressing that sum of squares in terms of basic ratios, we find that

$$\begin{aligned} SS_{error\ term} &= SS_{S/A} - SS_S \\ &= ([Y] - [A]) - ([S] - [T]) \\ &= [Y] - [A] - [S] + [T] \qquad (16\text{-}1) \end{aligned}$$

It is also possible to conceptualize this sum of squares as an *interaction*, which is how we will designate the error term in the remainder of the chapter. We pointed out already that the data matrix for the $(A \times S)$ design in Table 16-1 is in fact a factorial matrix, where the columns represent the levels of the independent variable (factor A) and the rows the "levels" of the subject "factor."[3] From our knowledge

[3] The only difference between the matrix in Table 16-1 and those associated with factorial designs we considered previously is that each cell of this matrix contains *one* observation rather than the sum of several.

of factorial designs, we would expect to subdivide the total sum of squares into sums of squares for two main effects (SS_A and SS_S) and an $A \times S$ interaction ($SS_{A \times S}$). There would be no "within-cell" variation, of course, since there is only one observation per cell in this matrix. We can now calculate the interaction sum of squares by subtracting the sums of squares for the two main effects from the total sum of squares:

$$\begin{aligned} SS_{A \times S} &= SS_T - SS_A - SS_S \\ &= ([Y] - [T]) - ([A] - [T]) - ([S] - [T]) \\ &= [Y] - [A] - [S] + [T] \end{aligned}$$

which is identical to Eq. (16-1).

Expressing the error term as an interaction sum of squares provides us with another way of understanding the nature of this new quantity: the $A \times S$ interaction represents the *unique manner* in which the different subjects respond to the treatment conditions. In other words, the error term consists of variability not attributable either to the treatment effects or to consistent individual differences.

Computational Formulas. The computational formulas for the overall analysis of variance are presented in Table 16-2. We have already discussed the formulas for the sums of squares. The formulas for the degrees of freedom require little comment, except for the error term, which reflects the form usually associated with an interaction. That is, the df for an interaction are generally specified as the product of the df's associated with the relevant main effects. In the present case,

$$\begin{aligned} df_{A \times S} &= (df_A)(df_S) \\ &= (a - 1)(s - 1) \end{aligned}$$

The remainder of the analysis is relatively straightforward. The treatment effects are assessed by evaluating the significance of an F ratio formed by dividing MS_A by $MS_{A \times S}$.

Table 16-2
ANOVA: Computational Formulas

Source	Basic Ratio*	df	Sum of Squares	MS	F
A	$[A] = \frac{\Sigma A^2}{s}$	$a - 1$	$[A] - [T]$	$\frac{SS_A}{df_A}$	$\frac{MS_A}{MS_{A \times S}}$
S	$[S] = \frac{\Sigma S^2}{a}$	$s - 1$	$[S] - [T]$	$\frac{SS_S}{df_S}$	
$A \times S$	$[Y] = \Sigma Y^2$	$(a - 1)(s - 1)$	$[Y] - [A] - [S] + [T]$	$\frac{MS_{A \times S}}{df_{A \times S}}$	
Total	$[T] = \frac{T^2}{(a)(s)}$	$(a)(s) - 1$	$[Y] - [T]$		

* Bracketed letters represent complete terms in computational formulas; a particular term is identified by the letter(s) appearing in the numerator.

A Numerical Example

For a numerical example, we return to the data we used to illustrate the analysis of the between-subjects single-factor experiment. In that experiment, subjects were randomly assigned to one of the three treatment conditions; lectures on physical science (a_1), social science (a_2), and history (a_3). Each condition was assigned 12 subjects, for a total of 36 subjects. Suppose instead that the experiment was an ($A \times S$) design in which we had only 12 subjects and each of the 12 subjects received all three lectures rather than only one. Consider the data presented in Table 16-3. The data matrix of Y scores (the AS matrix) displays the data by treatment and subject and provides all the information necessary to conduct an overall ANOVA.

Table 16-3
Numerical Example: *AS* Matrix

	Treatments			
	a_1	a_2	a_3	Sum
s_1	53	47	45	145
s_2	49	42	41	132
s_3	47	39	38	124
s_4	42	37	36	115
s_5	51	42	35	128
s_6	34	33	33	100
s_7	44	13	46	103
s_8	48	16	40	104
s_9	35	16	29	80
s_{10}	18	10	21	49
s_{11}	32	11	30	73
s_{12}	27	6	20	53
Sum	480	312	414	1206

Table 16-4
Summary of the Analysis

Source	Basic Ratio	Sum of Squares	*df*	*MS*	*F*
A	[*A*] = 41,595.00	[*A*] − [*T*] = 1194.00	2	597.00	12.30*
S	[*S*] = 43,932.67	[*S*] − [*T*] = 3531.67	11	321.06	
A × *S*	[*Y*] = 46,194	[*Y*] − [*A*] − [*S*] + [*T*] = 1067.33	22	48.52	
Total	[*T*] = 40,401.00	[*Y*] − [*T*] = 5793.00	35		

* $p < .01$.

We will assume that the order in which the lectures were given was systematically varied among the subjects, using some appropriate counterbalancing scheme. In the present case, one could arrange the three lectures in all six possible orders, namely,

1–2–3; 1–3–2; 2–1–3; 2–3–1; 3–1–2; and 3–2–1

and use each order twice, so that, for example, subjects 1 and 2 would receive order 1-2-3, subjects 3 and 4 would receive order 1-3-2, and so forth. However it is accomplished, the arrangement should guarantee that each lecture is presented equally often as the first, the second, or the third condition subjects encounter, in order to balance practice effects evenly among all three treatment conditions.[4]

Without comment, we will calculate the basic ratios needed for the ANOVA:

$$[T] = \frac{T^2}{(a)(s)} = \frac{1206^2}{(3)(12)} = 40{,}401.00$$

$$[A] = \frac{\Sigma A^2}{s} = \frac{480^2 + 312^2 + 414^2}{12} = 41{,}595.00$$

$$[S] = \frac{\Sigma S^2}{a} = \frac{145^2 + 132^2 + \cdots + 73^2 + 53^2}{3} = 43{,}932.67$$

$$[Y] = \Sigma Y^2 = 53^2 + 49^2 + \cdots + 30^2 + 20^2 = 46{,}194$$

The values of these basic ratios are entered in Table 16-4, where they are combined to produce the required sums of squares. The F of 12.30, which is evaluated with $df_{num.} = 2$ and $df_{denom.} = 22$, is significant at the $p < .01$ level.

[4] This experiment would probably also require the use of three different vocabulary tests. That is, each subject would receive a different vocabulary test following each lecture. Good experimental design would also require that each of the tests be used equally often with each of the three treatment conditions. Otherwise, it would not be possible to separate out the effects of the experimental treatments (lectures) from any differences in test difficulty.

We mentioned in Sec. 16.1 that when certain statistical assumptions are not met by the experimental data, a more conservative method of evaluating the significance of the obtained F may be necessary. One approach we suggested was to adopt a slightly more stringent significance level and to evaluate the F with this new critical value. To illustrate, we might consider setting $\alpha = .025$, for example, rather than equal to the usual .05. The critical value would now become 4.38 rather than the 3.44 normally appropriate when the statistical assumptions are met. In the present example, this new significance level does not change our decision; the F is still significant.[5]

16.3 THE OVERALL ANALYSIS: THE MRC APPROACH

There are several ways to approach the analysis of the ($A \times S$) design with MRC procedures. As you might suspect, the differences are in the coding used to represent the new critical source of variability. As we did when we discussed previous designs, we will emphasize contrast coding. The coding system needs to take into account the fact that "subjects" constitute a main effect. Thus, for within-subjects designs, the only new coding for us is the coding of "subjects"; otherwise, the coding system for treatment effects is the same as it was in the between-subjects design.

Coding the Main Effects

As we just stated, we will represent the variability of subjects with *contrast coding*. The number of vectors needed for this main effect is the number of subjects minus 1. With contrast coding, we easily construct a set of vectors whereby each vector in essence "compares" a subject systematically with each of the other subjects. The code matrix depicted in Table 16-5 indicates the results of this simple strategy. As you can see, the first subject has been used as the "comparison" subject, who is assigned a +1 in *all* the subject vectors (vectors 1 to 11); the remaining subjects

[5] An alternative approach is to use the so-called Geisser-Greenhouse correction (Geisser & Greenhouse, 1958), which gives us the appropriate critical value for the worst situation in which the assumptions are maximally violated. For this design, the correction consists of using $df_{num.} = 1$, rather than 2, and $df_{denom.} = s - 1 = 11$, rather than 22. The new critical value of F is now 4.84, and again, the observed F is significant. The Geisser-Greenhouse correction is only necessary when assumptions are maximally violated, however; the correction for in-between situations—between maximum violation and no violation—is more complicated (see Keppel, 1982, pp. 470–473).

Table 16-8
Vectors Representing the $A \times S$ Interaction

	Subject	Y	Vectors									
			A1S1	A1S2	⋯	A1S10	A1S11	A2S1	A2S2	⋯	A2S10	A2S11
a_1	1	53	1	1	⋯	1	1	1	1	⋯	1	1
	2	49	−1	0	⋯	0	0	−1	0	⋯	0	0
	3	47	0	−1	⋯	0	0	0	−1	⋯	0	0
	4	42	0	0	⋯	0	0	0	0	⋯	0	0
	5	51	0	0	⋯	0	0	0	0	⋯	0	0
	6	34	0	0	⋯	0	0	0	0	⋯	0	0
	7	44	0	0	⋯	0	0	0	0	⋯	0	0
	8	48	0	0	⋯	0	0	0	0	⋯	0	0
	9	35	0	0	⋯	0	0	0	0	⋯	0	0
	10	18	0	0	⋯	0	0	0	0	⋯	0	0
	11	32	0	0	⋯	−1	0	0	0	⋯	−1	0
	12	27	0	0	⋯	0	−1	0	0	⋯	0	−1
a_2	1	47	−1	−1	⋯	−1	−1	1	1	⋯	1	1
	2	42	1	0	⋯	0	0	−1	0	⋯	0	0
	3	39	0	1	⋯	0	0	0	−1	⋯	0	0
	4	37	0	0	⋯	0	0	0	0	⋯	0	0
	5	42	0	0	⋯	0	0	0	0	⋯	0	0
	6	33	0	0	⋯	0	0	0	0	⋯	0	0
	7	13	0	0	⋯	0	0	0	0	⋯	0	0
	8	16	0	0	⋯	0	0	0	0	⋯	0	0
	9	16	0	0	⋯	0	0	0	0	⋯	0	0
	10	10	0	0	⋯	0	0	0	0	⋯	0	0
	11	11	0	0	⋯	1	0	0	0	⋯	−1	0
	12	6	0	0	⋯	0	1	0	0	⋯	0	−1
a_3	1	45	0	0	⋯	0	0	−2	−2	⋯	−2	−2
	2	41	0	0	⋯	0	0	2	0	⋯	0	0
	3	38	0	0	⋯	0	0	0	2	⋯	0	0
	4	36	0	0	⋯	0	0	0	0	⋯	0	0
	5	35	0	0	⋯	0	0	0	0	⋯	0	0
	6	33	0	0	⋯	0	0	0	0	⋯	0	0
	7	46	0	0	⋯	0	0	0	0	⋯	0	0
	8	40	0	0	⋯	0	0	0	0	⋯	0	0
	9	29	0	0	⋯	0	0	0	0	⋯	0	0
	10	21	0	0	⋯	0	0	0	0	⋯	0	0
	11	30	0	0	⋯	0	0	0	0	⋯	2	0

In essence, the MRC strategy allows us at least two options for determining the error term:[7]

1. To calculate the error term directly by creating interaction vectors based on the vectors established for A and S
2. To calculate the error term indirectly by summing $R^2_{Y.A}$ and $R^2_{Y.S}$ to obtain $R^2_{Y.max.}$ and subtracting this sum from 1, as specified in Eq. (16-2)

16.4 COMPARISONS INVOLVING THE TREATMENT MEANS

As we have stated throughout this book, researchers usually design single-factor experiments with specific comparisons between the treatment conditions in mind. The same analytical procedures available for the analysis of a between-subjects design are available for a within-subjects design. The only complication is the determination of the *error term.* With the completely randomized design, the error term for the omnibus analysis ($MS_{S/A}$) is used as the denominator term of the F ratio for *any* comparison undertaken in the analysis. This procedure was justified by the assumption that population treatment variances are equal, which implies that the overall error term provides a perfectly suitable estimate of error variance present in any comparison conducted on the treatment means.[8]

With the ($A \times S$) design, on the other hand, there is no assurance that the error term from the omnibus analysis ($MS_{A \times S}$) can serve a similar function in the detailed analysis of an experiment. This certainty is lacking because the $A \times S$ interaction is actually an average of a number of *component interactions,* which may or may not be appropriately estimated by the overall error term. In many cases, they are not. The solution to this problem is conceptually simple, namely, to use as error terms $A \times S$ interactions that are each relevant to specific *single comparisons,* which we will call $A_{comp.} \times S$ interactions. As you will see, each of these error

[7] A third method of coding is frequently recommended for the analysis of within-subjects designs called *sum coding* or *criterion scaling* (see, for example, Edwards, 1979, pp. 120–123; Pedhazur, 1977). This ingenious method captures the entire source of subject variability with a *single vector,* in that the vector contains the sum of scores across all of the conditions for each subject. Unfortunately, this method of coding is *not* useful for conducting single-*df* comparisons, which, for most researchers, is the primary purpose of an experiment and its analysis.

[8] This statement is correct only when the homogeneity assumption is reasonably met by the data. With heterogeneous variances, special error terms are recommended (see Keppel,

terms is based *only* on the data contributing to the particular comparison under study.

Computational Procedures: ANOVA

A single-*df* comparison $A_{comp.}$ reflects the variability associated with the difference between two means. The error term for the mean square based on this difference reflects the degree to which individual subjects deviate from this average difference. This particular variation is represented by an *interaction*, $A_{comp.} \times S$, which, as we have already noted, is not necessarily estimated by the overall $A \times S$ interaction. We will consider the analysis of two comparisons, one a comparison between two means (physical science and social science) and the other a complex comparison (combined sciences and history).

The Computational Formulas. We begin with the comparison between physical science and social science. Consider the data arrangement in Table 16-9. The original *AS* matrix of *Y* scores is presented on the left. In the middle are the scores for each subject that will actually enter into this analysis, namely, a score obtained after the physical science lecture and a score obtained after the social science lecture. Consider this *comparison matrix* carefully. What we have is a matrix of scores that could be viewed as the results of a "miniature" ($A \times S$) design with only *two levels*. If in fact there were only two levels, we could calculate the standard sources of variance—an *A*, an *S*, and an $A \times S$ interaction—using the formulas from Sec. 16.2. The *F* test would consist of dividing the treatment mean square MS_A by the error term, $MS_{A \times S}$. This is exactly what we will accomplish in the analysis, except that the "treatment effect" is really the comparison effect ($MS_{A_{comp.}}$) and the "error term" is an interaction based only on the data involved in the comparison ($MS_{A_{comp.} \times S}$).

Table 16-9
Calculation of a Separate Error Term

	AS Matrix			Comparison Matrix		Calculations	
Subject	a_1	a_2	a_3	Physical Science	Social Science	Difference	*SS*
1	53	47	45	53	47	6	18.00
2	49	42	41	49	42	7	24.50
3	47	39	38	47	39	8	32.00
4	42	37	36	42	37	5	12.50
5	51	42	35	51	42	9	40.50
6	34	33	33	34	33	1	.50
7	44	13	46	44	13	31	480.50
8	48	16	40	48	16	32	512.00
9	35	16	29	35	16	19	180.50
10	18	10	21	18	10	8	32.00
11	32	11	30	32	11	21	220.50
12	27	6	20	27	6	21	220.50
			Mean:	40.00	26.00	14.00	

There are several ways to calculate the necessary sums of squares. The method we will illustrate focuses on the *difference* score for each subject. The differences are given in the right-hand portion of Table 16-9. For the first subject, the difference is 53 − 47 = 6; for the second subject, the difference is 49 − 42 = 7; and so on. The bottom row of the table gives the means for the two treatment conditions and the difference between them; that is,

$$\hat{\psi} = 40.00 - 26.00 = 14.00$$

The first step is to transform these various differences into sums of squares. The sum of squares for any given subject is given by

$$SS_{A_{comp.}\text{ for } s_j} = \frac{(\text{Diff.})^2}{\Sigma\,(c_i)^2} \qquad (16\text{-}3)$$

(In a moment, we will eventually combine the sums of squares from all the subjects.) Next, we calculate the comparison sum of squares $SS_{A_{comp.}}$, using the formula originally presented in Chap. 11:

$$SS_{A_{comp}} = \frac{(s)\,(\hat{\psi})^2}{\Sigma\,(c_i)^2} \qquad (16\text{-}4)$$

The interaction sum of squares $SS_{A_{comp.} \times S}$ is obtained by subtracting the comparison sum of squares $SS_{A_{comp.}}$ from the sum of the subject sums of squares. That is,

$$SS_{A_{comp.} \times S} = \sum \frac{(\text{Diff.})^2}{\Sigma\,(c_i)^2} - SS_{A_{comp.}} \qquad (16\text{-}5)$$

(Although we will not demonstrate it here, this sum of squares is identical to an "$SS_{A \times S}$" obtained by treating the comparison matrix as an actual within-subjects design.)

Before we turn to the numerical example, let us look at the operations specified in Eq. (16-5). Consider the first quantity on the right side of the equation. This is a composite sum of squares that combines the treatment sums of squares for the individual subjects. This quantity reflects the difference between the two treatments and the *unique way* in which each subject responds to the two treatments (the interaction). To make this sum useful as an error term, we must remove the *systematic source* of variability—the treatment difference—from the composite sum of squares. This is precisely what we accomplished by subtracting $SS_{A_{comp.}}$ in Eq. (16-5).

A Numerical Example. We will illustrate the calculations with the data in Table 16-9. Using Eq. (16-3), we find the comparison sum of squares for the first subject to be:

$$SS_{A_{comp.}\text{ for } s_1} = \frac{(6)^2}{(+1)^2 + (-1)^2 + (0)^2} = 18.00$$

The sums of squares for all 12 subjects are given in the final column of the table. The sum of these individual sums of squares is

$$\begin{aligned}\Sigma SS_{A_{comp.}\text{ for } s_j} &= 18.00 + 24.50 + \cdots + 220.50 + 220.50\\ &= 1774.00\end{aligned}$$

For the comparison sum of squares, we substitute the difference between the two means in Eq. (16-4) and find

$$SS_{A_{comp.}} = \frac{(12)(14.00)^2}{(+1)^2 + (-1)^2 + (0)^2} = 1176.00$$

Finally, from Eq. (16-5), we obtain

$$SS_{A_{comp.} \times S} = 1774.00 - 1176.00 = 598.00$$

The formula for the F is given by Eq. (16-6):

$$F_{comp.} = \frac{MS_{A_{comp.}}}{MS_{A_{comp}} \times S} \qquad (16\text{-}6)$$

Since there is 1 df for the comparison, $MS_{A_{comp}} = 1176.00$. The degrees of freedom associated with the error term are given by Eq. (16-7):

$$\begin{aligned}df_{A_{comp.} \times S} &= (df_{A_{comp}})(df_S)\\ &= (1)(s-1) = s-1 \qquad (16\text{-}7)\end{aligned}$$

For the present example, $df_{A_{comp.} \times S} = 12 - 1 = 11$, and $MS_{A_{comp.} \times S} = 598.00/11 = 54.36$. Substituting in Eq. (16-6), we find

$$F_{comp.} = \frac{1176.00}{54.36} = 21.63$$

which is significant. (As a reminder, this F is evaluated with $df_{num.} = 1$ and $df_{denom.} = 11$.) You should note that the error term for this comparison does *not* equal the error term for the omnibus F test ($MS_{A \times S} = 48.52$).

A Second Numerical Example. For a second example, we will compare the average of the two science conditions with the history condition. Table 16-10 presents the relevant comparison matrix. Entries in the first column consist of each subject's average vocabulary score obtained following the two science lectures; entries in the second column consist of each subject's score obtained following the history lecture. The difference scores and sums of squares based on these differences are found in the last two columns of the table. For the first subject, the average science score is $(53 + 47)/2 = 50.00$ and the history score is 45.00—a difference of 5.00 words. If we translate this difference into a sum of squares, we find

$$SS_{A_{comp.}\text{ for } s_1} = \frac{(5.00)^2}{(+\frac{1}{2})^2 + (+\frac{1}{2})^2 + (-1)^2} = 16.67$$

The sum of the sums of squares for all subjects is:

$$\begin{aligned}\Sigma SS_{A_{comp.}\text{ for } s_j} &= 16.67 + 13.50 + \cdots + 48.17 + 8.17\\ &= 487.37\end{aligned}$$

Table 16-10
Calculation of a Separate Error Term

	Comparison Matrix		Calculations	
Subject	Combined Science	History	Difference	SS
1	50.00	45	5.00	16.67
2	45.50	41	4.50	13.50
3	43.00	38	5.00	16.67
4	39.50	36	3.50	8.17
5	46.50	35	11.50	88.17
6	33.50	33	.50	.17
7	28.50	46	−17.50	204.17
8	32.00	40	−8.00	42.67
9	25.50	29	−3.50	8.17
10	14.00	21	−7.00	32.67
11	21.50	30	−8.50	48.17
12	16.50	20	−3.50	8.17
Mean:	33.00	34.50	−1.50	

Using the data at the bottom of Table 16-10, we find:

$$SS_{A_{comp.}} = \frac{(12)(-1.50)^2}{(+\frac{1}{2})^2 + (+\frac{1}{2})^2 + (-1)^2} = 18.00$$

The sum of squares for the error term is

$$SS_{A_{comp.} \times S} = 487.37 - 18.00 = 469.37$$

Continuing with the calculations, we find $MS_{A_{comp.}} = 18.00$ and $MS_{A_{comp.} \times S} = 469.37/11 = 42.67$. The $F_{comp.}$ is not significant ($18.00/42.67 = .42$).

Comment. It is interesting that the two sums of squares we obtained for error terms for the two comparisons add up to the sum of squares for the error term from the overall analysis. That is,

$$598.00 + 469.37 = 1067.37$$

which, except for rounding error, equals $SS_{A \times S}$ (1067.33). This has occurred only because the two comparisons were *orthogonal*, a property that extended to the two error terms, which is why the two sums of squares totaled $SS_{A \times S}$.

You may have noticed that the two error terms in our hypothetical experiment differed by a relatively small amount (54.36 versus 42.67). In actual experiments, however, sizable differences do occur; and even small differences, such as these, can affect the outcome of a statistical test. For these reasons, we recommend that you use separate error terms for the different comparisons of interest unless there is convincing evidence that such a procedure is not necessary.

Computational Procedures: MRC

The analysis of single-*df* comparisons with MRC is a relatively simple matter and is most easily conducted by adopting a strategy of assuming that an omnibus analysis is necessary (though in fact it is not), coding for comparisons and subjects, and creating vectors for interactions of treatments with subjects. As you know, the variation associated with the comparison itself is reflected in the zero-order correlation between Y and the appropriately coded comparison vector. As we will show, the variation associated with an $A_{comp.} \times S$ interaction is reflected in the subset of the *interaction vectors* formed by cross-multiplying the particular comparison vector with all the individual subject vectors.

Consider the vectors in Table 16-11. Vector 1 ($A1$) specifies the comparison between the two science conditions. The vectors in columns 2 through 12 are the interaction vectors found by cross-multiplying the numerical values of vector $A1$ with corresponding values from all the subject vectors presented in Table 16-6.

Table 16-11
Vectors Representing a Single-*df* Comparison

			Vectors											
	Subject	Y	(1) A1	(2) A1S1	(3) A1S2	(4) A1S3	(5) A1S4	(6) A1S5	(7) A1S6	(8) A1S7	(9) A1S8	(10) A1S9	(11) A1S10	(12) A1S11
a_1	1	53	1	1	1	1	1	1	1	1	1	1	1	1
	2	49	1	−1	0	0	0	0	0	0	0	0	0	0
	3	47	1	0	−1	0	0	0	0	0	0	0	0	0
	4	42	1	0	0	−1	0	0	0	0	0	0	0	0
	5	51	1	0	0	0	−1	0	0	0	0	0	0	0
	6	34	1	0	0	0	0	−1	0	0	0	0	0	0
	7	44	1	0	0	0	0	0	−1	0	0	0	0	0
	8	48	1	0	0	0	0	0	0	−1	0	0	0	0
	9	35	1	0	0	0	0	0	0	0	−1	0	0	0
	10	18	1	0	0	0	0	0	0	0	0	−1	0	0
	11	32	1	0	0	0	0	0	0	0	0	0	−1	0
	12	27	1	0	0	0	0	0	0	0	0	0	0	−1
a_2	1	47	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
	2	42	−1	1	0	0	0	0	0	0	0	0	0	0
	3	39	−1	0	1	0	0	0	0	0	0	0	0	0
	4	37	−1	0	0	1	0	0	0	0	0	0	0	0
	5	42	−1	0	0	0	1	0	0	0	0	0	0	0
	6	33	−1	0	0	0	0	1	0	0	0	0	0	0
	7	13	−1	0	0	0	0	0	1	0	0	0	0	0
	8	16	−1	0	0	0	0	0	0	1	0	0	0	0
	9	16	−1	0	0	0	0	0	0	0	1	0	0	0
	10	10	−1	0	0	0	0	0	0	0	0	1	0	0
	11	11	−1	0	0	0	0	0	0	0	0	0	1	0
	12	6	−1	0	0	0	0	0	0	0	0	0	0	1

19

Higher-Order Factorial Designs

. . .

19.1 THE COMPLETELY RANDOMIZED THREE-FACTOR DESIGN
- The Design
- The $A \times B \times C$ Interaction
- Sources of Variance
- Using ANOVA to Perform the Overall Analysis
- Using MRC to Perform the Overall Analysis

19.2 DETAILED ANALYSIS OF THE $A \times B \times C$ INTERACTION
- Analysis of Simple Effects
- Analysis of Interaction Comparisons

19.3 HIGHER-ORDER WITHIN-SUBJECTS DESIGNS
- "Pure" Between-*S* Designs
- "Pure" Within-*S* Designs
- The Mixed Two-Factor Design: A Review
- Mixed Three-Factor Designs

19.4 USING ANOVA AND MRC TO ANALYZE WITHIN-SUBJECTS DESIGNS
- The ANOVA Approach
- The MRC Approach

19.5 SUMMARY

19.6 EXERCISES

This chapter is included to give you a glimpse at the complexity of higher-order factorial designs and an appreciation of their explanatory potential. Some of you may find this material too abstract and too advanced at the present stage in your training.[1] We suggest that you view this chapter as a sample of the capabilities of experimental design, rather than a complete treatment of the subject. Our goal is to indicate how the principles underlying simpler experimental designs covered in the preceding chapters apply to more complex designs. We hope that this chapter will give you a perspective from which you can design more satisfying and ambitious projects in your future research efforts.

Factorial designs are easily expanded to form higher-order factorial designs—designs that incorporate more than two independent variables. The higher-order factorial designs we will discuss in this chapter all have the same defining property: they include all possible combinations of the levels of the factors in the experiment. A 2 × 3 × 2 design, for example, includes a total of $(2)(3)(2) = 12$ treatment conditions formed by crossing the two levels of factor A with the three levels of factor B and then crossing the resulting six combinations with the two levels of factor C. Designs in which we explore all possible combinations are called **completely crossed factorials.**[2]

With an increase in the number of independent variables comes a marked increase in the amount of information that you can obtain. That is, while the overall analysis of higher-order factorials still consists of the examination of main effects and interactions, you will find that they offer more main effects and interactions to study. Moreover, the *complexity* of the interactions increases as well. In a moment, we will consider the nature of the information that higher-order factorial designs provide.

This chapter can only scratch the surface of the topic. As we have indicated already, our intent is to show how the principles we considered in our discussions of the analysis of one- and two-factor designs *generalize* to the analysis of higher-order designs. At some point you will probably have to consult more comprehensive discussions of this material, but at least you will have some appreciation of how the analyses of complex designs are derived from the analyses of simpler ones.

[1] Note to instructors: The material in this chapter may be omitted without affecting your students' understanding of the remaining chapters.

[2] Incomplete factorials, in which *not* all possible combinations are included, are relatively uncommon in the behavioral sciences. For a discussion of these designs, see Kirk (1982, pp. 489–710) and Winer (1971, pp. 604–684).

19.1 THE COMPLETELY RANDOMIZED THREE-FACTOR DESIGN

We will consider briefly the design and analysis of the three-factor design. As we have indicated, the three-factor design is made up of all possible combinations of the levels of three independent variables. For this discussion, we are assuming that subjects are randomly assigned, in equal numbers (s), to the $(a)(b)(c)$ different treatment conditions. Although repeated measures can easily be introduced into this design, they complicate the evaluation process considerably, because they call for a variety of error terms to conduct the different statistical tests. In comparison, the evaluation process for the completely randomized three-factor design is relatively simple, since only one error term is needed to conduct these same tests. It is important to note, however, that all three-factor designs supply the same type of information, which means that we can focus on the information provided by these designs without worrying at this time about the complications created when repeated measures are introduced. We discuss within-subjects designs in Sec. 19.3.

The Design

The three-factor design is a natural outgrowth of the two-factor design. To illustrate, we will start with the 3 × 2 factorial we used to introduce the $A \times B$ design, an experiment consisting of the three types of lectures (physical science, social science, and history) and two methods of presentation (computer and standard). Suppose we add a third independent variable (factor C)—a developmental variable, age—consisting of $c = 2$ levels, fifth-grade and eighth-grade children. This design is presented in Table 19-1. You will note that the original $A \times B$ design is represented *twice* in this three-factor design, once in conjunction with level c_1 (fifth grade) and once with level c_2 (eighth grade). Taken as a whole, the complete design is made up of all possible combinations of the levels of the three independent variables.

The $A \times B \times C$ Interaction

The one new concept we introduced when we first discussed the two-factor design was, of course, the $A \times B$ interaction. No new concepts are introduced with higher-order designs, although the interactions may be of greater complexity. We will illustrate this point with the $A \times B \times C$ interaction.

All interactions may be defined in terms of *simple effects*. As a reminder from the two-factor case,

> **An $A \times B$ interaction is present when the simple effects of one of the independent variables are not the same at all levels of the second independent variable.**

Table 19-1
An Example of a Three-Factor Design

	Fifth-Grade Students (c_1)				Eighth-Grade Students (c_2)		
	Physical Science (a_1)	Social Science (a_2)	History (a_3)		Physical Science (a_1)	Social Science (a_2)	History (a_3)
Computer (b_1)				Computer (b_1)			
Standard (b_2)				Standard (b_2)			

The $A \times B \times C$ interaction is defined in an analogous fashion except that we focus on simple *interactions* rather than on simple *effects*. A **simple interaction** is the interaction of two of the independent variables with the third variable *held constant*. Translated to the present context, the presence of an $A \times B \times C$ interaction would mean that the interaction of factors A and B for the fifth-grade students (level c_1) is different from the corresponding interaction for the eighth-grade students (level c_2).[3] In words, then,

> **An $A \times B \times C$ interaction is present when the simple interactions between two of the independent variables are not the same at all levels of the third.**

We can often comprehend interactions more easily if we plot the means on a graph. Consider first one possible outcome of this experiment, which is presented in the upper half of Fig. 19-1. Compare the simple interaction on the left (fifth grade) with the simple interaction on the right (eighth grade). Remember that if the two seem to be different, an $A \times B \times C$ interaction may be present, whereas if they appear to be the same, there is probably no interaction present. Of course, we would eventually base our judgment on the outcome of an appropriate statistical test. For the moment, however, let us just examine the data informally, with an eye toward understanding the concept. The graph on the left suggests that there is an interaction between lectures and methods for the fifth-grade children, while the graph on the right reflects no interaction whatsoever. Since the two simple interactions are *not the same*—we find a sizable interaction on the left and no interaction on the right—we would conclude that an $A \times B \times C$ interaction is present. The next step would be a statistical test that assesses this difference.

In contrast, consider the outcome depicted in the lower portion of Fig. 19-1. In this case, apparently, the same interaction found with the fifth-grade students is also found with the eighth-grade students. This suggests that the two simple interactions are roughly alike and that a statistical test would be likely to reveal that an $A \times B \times C$ interaction is not present.

Sources of Variance

The standard analysis of the three-way factorial examines three types of treatment effects. One of these is the $A \times B \times C$ interaction, of course, which is based on all the individual means of the different treatment conditions. The other two represent

[3] The $A \times B \times C$ interaction is defined in terms of the simple interactions created by the combination of any two of the three independent variables. We chose the simple $A \times B$ interaction. We could just as easily have chosen the simple $A \times C$ interaction at the two levels of factor B or the simple $B \times C$ interaction at the three levels of factor A.

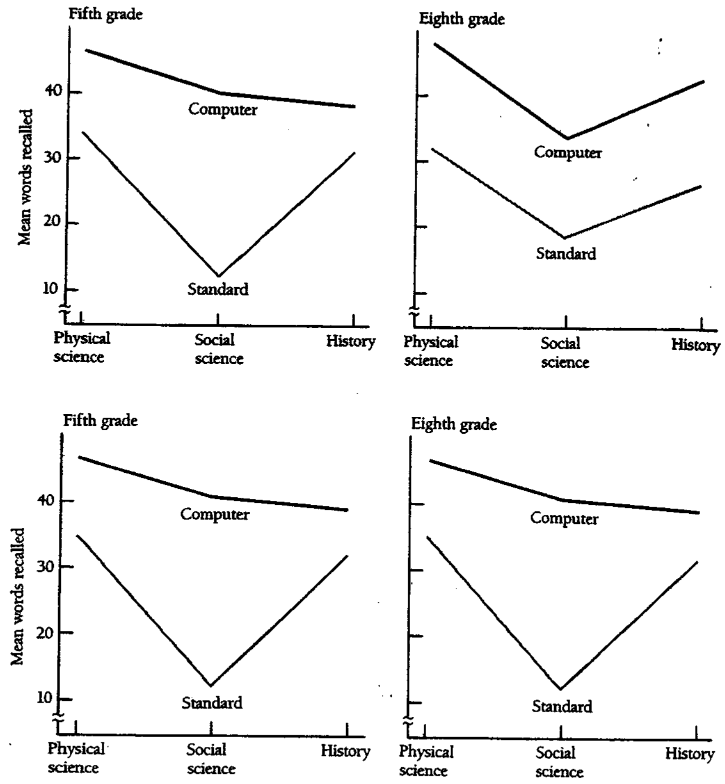


Figure 19-1 Two possible outcomes of the three-way factorial. The upper graph displays an $A \times B \times C$ interaction, while the lower graph displays no $A \times B \times C$ interaction.

sources created by averaging or collapsing over some of the treatment conditions. More specifically, we could disregard factor C completely by combining the data from the fifth- and eighth-grade students. This would leave us with what amounts to a two-factor design and an $A \times B$ interaction. Alternatively, we could combine the data from the two levels of factor B (computer and standard presentation) to create an $A \times C$ design and an $A \times C$ interaction and consequently disregard mode of presentation; or we could combine the data from the three levels of factor A (physical science, social science, and history) to create a $B \times C$ design and a $B \times C$ interaction, ignoring type of lecture. Finally, we could examine any of three *main effects*—A, B, and C—by pooling the data over the levels of the other two independent variables. The A main effect, for example, is based on the overall means for the three lectures obtained by averaging the data for all levels of factors B and C.

The interpretation of the results of this analysis generally begins with a test of the $A \times B \times C$ interaction, which we have indicated at the top of Fig. 19-2. A significant three-way interaction means that any effects based on data that are collapsed over the levels of one or two independent variables may provide a distorted picture of the influence of the three factors on the dependent variable. Consequently, the next step is to turn to special analyses that help to identify the sources of the significant interaction. We will discuss these analyses in Sec. 19.2. On the other hand, a nonsignificant interaction indicates that it is "safe" to continue the standard analysis by examining the data with *one* of the factors removed or disregarded. At this point, then, the analysis would focus on all possible interactions between two of the factors—$A \times B$, $A \times C$, and $B \times C$—just as if they had been produced from actual two-factor designs. This step is also shown in Fig. 19-2.

From now on, the analysis should be familiar. We test each of the interactions for significance. If an interaction is significant, we try to discover the differences responsible for it and pay little attention to the main effects. On the other hand, a main effect is of interest for a factor that is not involved in a significant two-way interaction. For example, if only the $A \times B$ interaction is significant, we can safely examine the C main effect because factor C is not involved in an interaction. If

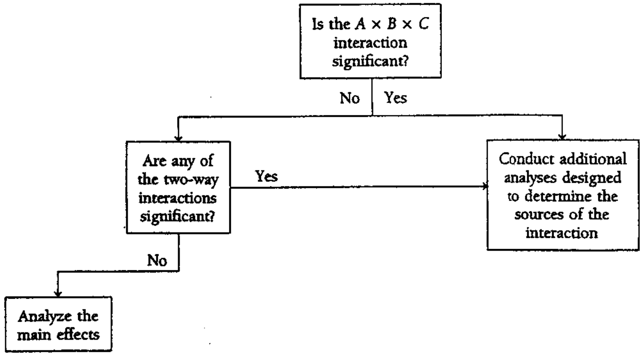

Is the A × B × C interaction significant?
No
Yes
Are any of the two-way interactions significant?
Yes
Conduct additional analyses designed to determine the sources of the interaction
No
Analyze the main effects

Figure 19-2 Analysis of a three-factor design.

another interaction is also significant, either $A \times C$ or $B \times C$, none of the main effects are left uncontaminated by interaction.[4]

Using ANOVA to Perform the Overall Analysis

The overall analysis represents a simple extension of the operations we followed in the analysis of single-factor and two-factor designs. We calculate sums of squares from basic ratios that are based on the various totals and subtotals we obtain when the data are collapsed over subjects and over the levels of the three independent variables. Degrees of freedom for the three main effects are equal to the number of levels less 1, while the *df* for interactions are found by multiplying the *df*'s associated with the factors specified by the interaction. The *df* for the $B \times C$ interaction, for example, are

$$df_{B \times C} = (df_B)(df_C) = (b-1)(c-1)$$

while the *df* for the $A \times B \times C$ interaction are

$$df_{A \times B \times C} = (df_A)(df_B)(df_C) = (a-1)(b-1)(c-1)$$

The mean squares are calculated in the usual fashion by dividing *SS*'s by the appropriate *df*'s. The error term for this analysis is based on the within-group variances obtained for the separate treatment conditions, which are pooled and averaged over all the $(a)(b)(c)$ treatment conditions.

Using MRC to Perform the Overall Analysis

The key to the MRC analysis, as in other designs, is the creation of vectors designed to represent the three main effects. As you know, the number of vectors required is, in essence, specified by the *df* associated with each source of variability isolated in the analysis. To illustrate, we will start with the main effects. In the present example, we would need two vectors to capture the *A* main effect (*A*1 and *A*2), one to capture the *B* main effect (*B*1), and one to capture the *C* main effect (*C*1). In general, vectors representing interactions are formed by cross-multiplying the vectors of the relevant main effects. For the $A \times C$ interaction, for example, the cross multiplication would involve the two vectors associated with the *A* main effect and the single vector associated with the *C* main effect, creating *A*1*C*1 and *A*2*C*1. Together, these two vectors capture the variation due to the $A \times C$ interaction. Similarly, we represent the $A \times B$ interaction with two interaction vectors, *A*1*B*1 and *A*2*B*1, and the $B \times C$ interaction with one interaction vector, *B*1*C*1.

[4] The logic of this analysis strategy is covered more fully in Keppel (1982, pp. 295–297).

Finally, we represent the $A \times B \times C$ interaction with interaction vectors created by the cross multiplication of all possible combinations of the three main-effects vectors. There are two such interaction vectors in this example,

*A*1*B*1*C*1 and *A*2*B*1*C*1

(How we multiply these main-effect vectors will be discussed in Sec. 19.2.) The squared multiple correlation coefficients involving the dependent variable *Y* and the relevant sets of vectors represent the proportions of variability associated with the different factorial effects. The *df* for each correlation are determined by the number of vectors required to define the particular effect. The residual variation is obtained by subtracting $R^2_{Y.max.}$, which is the squared multiple correlation coefficient between *Y* and all the factorial vectors, from 1. That is,

$$R^2_{Y.max.} = R^2_{Y.A} + R^2_{Y.B} + R^2_{Y.C} + R^2_{Y.A \times B} + R^2_{Y.A \times C} + R^2_{Y.B \times C} + R^2_{Y.A \times B \times C}$$
$$R^2_{residual} = 1 - R^2_{Y.max}$$

19.2 DETAILED ANALYSIS OF THE $A \times B \times C$ INTERACTION

Two general techniques are available for determining the factors responsible for a significant $A \times B \times C$ interaction. These are the *analysis of simple effects* and the *analysis of interaction comparisons*—both variations of the same techniques we used to analyze the $A \times B$ interaction in Chaps. 14 and 15. We are not able to cover these procedures in detail, but we will emphasize their nature and form. A comprehensive discussion of this material may be found in Keppel (1982, Chap. 14).

Analysis of Simple Effects

All interactions may be expressed as differences in simple effects. For the $A \times B$ interaction, for example, the differences of interest to us are in the effects of one of the independent variables at different levels of the other. For the complex or higher-order $A \times B \times C$ interaction, on the other hand, we are interested in the differences in the *interaction* of two of the factors at different levels of the third. Once we detect an interaction, an obvious next step is to examine the simple effects themselves in an attempt to establish its exact nature.

The simple effects of any interaction are revealed by subdividing the original factorial design into a set of less complex designs, each of which defines a different simple effect. You will recall from Chap. 14, for example, that we can uncover the simple effects of the $A \times B$ interaction by analyzing a set of component *single*-factor designs in which we vary one of the independent variables while holding the other constant. We might examine the effects of factor *A* at level b_1, at level b_2, and so on, or the effects of factor B at level a_1, at a_2, and so on. By transforming a more complex experiment (in this case, the two-factor design) into a set of less complex experiments (component single-factor designs), we are able to discover the ways in which the simple effects differ from one another.

We follow this same general procedure with the three-factor design and the analysis of the $A \times B \times C$ interaction. We transform the more complex experiment (the three-factor design) into a set of less complex experiments, which in this case consist of component *two*-factor designs. These latter designs involve the manipulation of two of the independent variables with the third held constant. There are three possibilities:

An $A \times B$ design at level c_1, at level c_2, and so on
An $A \times C$ design at level b_1, at level b_2, and so on
A $B \times C$ design at level a_1, at level a_2, and so on

A significant $A \times B \times C$ interaction means that the simple interactions are not the same, and this is true for any one of these sets of component factorial designs.

Most researchers have a preferred way of expressing the $A \times B \times C$ interaction and thus will usually examine only the sets in which they are most interested if the interaction proves to be significant. Assuming that the $A \times B \times C$ interaction reflected in the data presented in the upper portion of Fig. 19-1 is significant, we would probably look at the simple interaction of lectures (factor A) and presentation (factor B), first for the fifth-grade students (level c_1—the display on the left)—and then for the eighth-grade students (level c_2—the display on the right). We would choose this alternative because of the way we conceptualized the experiment—the joint manipulation of lectures and presentation at two different age levels.

An analysis of these simple interactions would probably reveal a significant interaction of lectures and presentation for the younger students, but no interaction for the older ones. If this were the case, we would pay little or no attention to the simple interaction for the older students and would concentrate our efforts on analyzing the simple interaction for the younger students. Thus, we would probably consider additional analyses in an attempt to identify the factors contributing to the significant interaction of lectures and methods of presentation for the fifth-grade students. At this point, the analysis exactly resembles the analysis of the simple effects of a significant interaction in an actual *two*-factor design. Thus, we might look at the effects of the three lectures first with the computer presentation and then with the standard presentation. If either of these effects is significant, we could examine meaningful single-*df* comparisons, such as the difference between the two science conditions and the difference between the combined science conditions and the history condition.

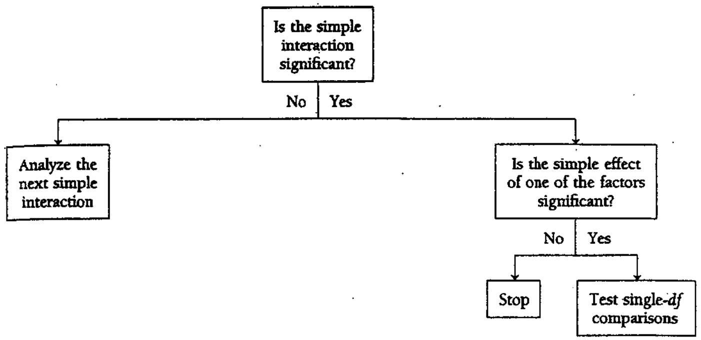

Is the simple interaction significant?
No
Yes
Analyze the next simple interaction
Is the simple effect of one of the factors significant?
No
Yes
Stop
Test single-*df* comparisons

Figure 19-3 Analysis of the three-way interaction.

While this process may sound complicated, it does represent a consistent pattern in which the discovery of a significant higher-order effect is followed by the analysis of a relevant simple effect. As a summary of this approach to the analysis of the three-way interaction, we could say that

A significant $A \times B \times C$ interaction is followed by an analysis of simple interactions, for example, the $A \times B$ interaction for fifth-grade students and for eighth-grade students.

A significant simple $A \times B$ interaction is then followed by the analysis of the simple effects of this interaction, for example, the effects of the different lectures (A) for the fifth-grade students under the two methods of presentation.

A significant simple effect of factor A is then followed by an analysis of meaningful single-*df* comparisons.

These steps are also diagramed in Fig. 19-3.

The ANOVA Approach. As you have seen, the simple interaction effects of the $A \times B \times C$ interaction are interactions obtained from component two-factor experiments. What this means is that you may calculate the sums of squares required for the analysis by isolating the data matrix of interest and then applying formulas appropriate for an actual two-factor experiment. From this point on, you will be on familiar ground; you can take advantage of the formulas we covered in Sec. 14.2. The only change is in the error term for the analysis. For all the simple interactions, the error term would come from the overall analysis of the three-factor design and would be $MS_{S/ABC}$, which is based on within-group variability pooled over the (*a*) (*b*) (*c*) treatment groups.

The MRC Approach. You may easily accomplish the analysis of simple effects with MRC by creating special vectors that reflect the desired simple effects. For the simple $A \times B$ interaction at level c_1, for example, you would start with the vectors defining the $A \times B$ interaction and then modify them for the analysis by assigning 0s to all observations at all levels of factor C not involved in the analysis (level c_2 in this case). You would create the vectors for the $A \times B$ interaction at level c_2 by starting with the same original interaction vectors, but modifying them by assigning 0s to all observations at level c_1. The error term comes from the overall analysis of the entire factorial design, Mean $R^2_{Y.S/ABC}$. For details concerning the analysis of the simple effects, see the discussion in Sec. 14.3.

Analysis of Interaction Comparisons

A second major way of analyzing interaction consists of creating a number of smaller factorial designs by transforming one or more of the independent variables into a form that reflects single-*df* comparisons. This transformation is possible, of course, only when a factor consists of more than two levels. For the example in Table 19-1, only factor *A*—with the levels consisting of three different lectures—qualifies as a candidate; the other two factors each consist of two levels.

Table 19-2 illustrates two component factorial designs created from single-*df* comparisons involving factor *A*. The analysis in the top half of the table is a three-factor design consisting of a comparison between the two science conditions ($A_{comp.\ 1}$) and the other two factors (method of presentation and grade level). The analysis in the bottom half is a three-factor design consisting of a comparison between the combined science conditions and the history condition ($A_{comp.\ 2}$) and the other two factors. The focus of both of these analyses would be the $A_{comp.} \times B \times C$ interactions.

It is important to appreciate the value of this type of analysis. A significant $A \times B \times C$ interaction found in the overall analysis does not indicate what aspects of factor *A* are critical. In contrast, the two component factorials provide a more insightful view of the interaction of the three factors. A significant $A_{comp.} \times B \times C$ interaction in the first case would bring attention to the differences between the two science lectures; a nonsignificant interaction would imply that any differences between the two science lectures have little effect on the interaction of the three factors. By the same token, the second case focuses attention on differences between science and history lectures, with no differentiation made between the two science conditions.

Table 19-2
Two Examples of Interaction Contrasts

	Fifth Grade			Eighth Grade	
	Physical Science	Social Science		Physical Science	Social Science
Computer			Computer		
Standard			Standard		

	Fifth Grade			Eighth Grade	
	Combined Science	History		Combined Science	History
Computer			Computer		
Standard			Standard		

Three-factor interaction comparisons associated with a single *df* are called, like their two-factor counterpart, **interaction contrasts**. They are produced from what is conceptually equivalent to a 2 × 2 × 2 design. Such interaction contrasts are valuable because the three-way interaction contrast (2 × 2 × 2), which is associated with 1 *df*, takes the analysis to the basic level since it cannot be divided into additional component factorials. Both of the examples in Table 19-2 are interaction contrasts.

Other types of interaction comparisons are possible in which some, but not all of the factors are represented by single-*df* comparisons. For example, suppose we have a 3 × 3 × 3 factorial. If only one of the factors (we will choose *A* arbitrarily) lends itself to analytical treatment, the analysis would consist of a number of 2 × 3 × 3 component factorials of the form $A_{comp.} \times B \times C$. If two of the factors (*A* and *B*) are transformed into single-*df* comparisons, the analysis will consist of a number of 2 × 2 × 3 component factorials of the form $A_{comp.} \times B_{comp.} \times C$. These various possibilities are discussed in detail by Keppel (1982, pp. 315–320).

The ANOVA Approach. We will concentrate our discussion on the analysis of interaction contrasts. The only new operation is the calculation of the contrast itself ($\hat{\psi}_{A \times B \times C}$), which may be obtained directly from a systematic layout of the means. Suppose we were interested in the three-way interaction of the two science conditions, the two methods of presentation, and the two grade levels. We would start by forming a 2 × 2 × 2 matrix corresponding to this interaction contrast. This we have accomplished in Table 19-3, using the data appearing in the upper portion of Fig. 19-1. You will recall that a three-way interaction is defined as the presence of differences in the interaction of two of the independent variables over the levels of the third independent variable. All that we need to do, then, is to compare the simple interaction of two of these factors at the two levels of the third factor.

The 2 × 2 matrix on the left in Table 19-3 provides information concerning the interaction of the two science lectures with the two methods of presentation for the fifth-grade students, while the matrix on the right provides corresponding information for the eighth-grade students. A value for each of the two interactions may be obtained simply by calculating the difference between the two means in each row and then subtracting the differences for each simple interaction. In this example,

$$\hat{\psi}_{A \times B \text{ at level } c_1} = 6.00 - 22.00 = -16.00$$

$$\hat{\psi}_{A \times B \text{ at level } c_2} = 14.00 - 14.00 = 0.00$$

The value of -16.00 for the interaction at c_1 indicates that there is an interaction between the two lectures and the two methods of presentation for the younger children, while the value of 0.00 indicates the complete absence of such an interaction for the older children. (A value of zero is highly unlikely, of course, because of the inevitable operation of chance factors.) The fact that the two "lecture" × "method" interactions are different (-16.00 versus 0.00) means that a three-way interaction is present. The value of this interaction contrast is

$$\hat{\psi}_{A \times B \times C} = (\hat{\psi}_{A \times B \text{ at level } c_1}) - (\hat{\psi}_{A \times B \text{ at level } c_2})$$
$$= (-16.00) - (0.00) = -16.00$$

The sum of squares associated with this value is obtained by substituting in a familiar formula, namely,

$$SS_{A_{comp.} \times B_{comp.} \times C_{comp.}} = \frac{(s)(\hat{\psi}_{A \times B \times C})^2}{[\Sigma\,(c_i)^2][\Sigma\,(c_j)^2][\Sigma\,(c_k)^2]}$$

where s = the sample size

$\hat{\psi}_{A \times B \times C}$ = the interaction contrast

c_i = the coefficients associated with factor A $(1, -1, 0)$

c_j = the coefficients associated with factor B $(1, -1)$

c_k = the coefficients associated with factor C $(1, -1)$

This sum of squares is based on 1 *df*, and the error term comes from the overall analysis.

Table 19-3
Calculating an Interaction Contrast

	Fifth Grade		
	Physical Science	Social Science	Difference
Computer	46.00	40.00	6.00
Standard	34.00	12.00	22.00

	Eighth Grade		
	Physical Science	Social Science	Difference
Computer	47.83	33.83	14.00
Standard	32.17	18.17	14.00

In all completely randomized designs, the within-groups error term is used to assess the significance of all treatment effects extracted in the analysis.[5] As you will see shortly, the same sort of error term will also be appropriate in *mixed* designs when the treatment effects are based entirely on between-*S* differences.

"Pure" Within-*S* Designs

There is a general principle that lies behind the error terms required for "pure" within-*S* designs. This rule may be stated as follows:

> **The error term for evaluating any source of variability in a "pure" within-*S* design is based on the interaction between the factor or factors contained in the source and subjects.**

Let us see how this rule functions in several common examples of "pure" within-*S* designs.

Single-Factor Within-*S* Design. You have already seen in Chap. 16 how this rule applies to the analysis of the single-factor within-*S* design. As a reminder, the error term for evaluating the overall effects of factor *A* is based on the interaction between factor *A* and subjects—the *A* × *S* interaction. You will also recall that the evaluation of single-*df* comparisons generally requires specialized error terms, unique to each comparison. If we substitute *comparison* for *source*, the rule applies to the analysis of these single-*df* comparisons as well. That is,

> **The error term for evaluating a single-*df* comparison is based on the interaction between the comparison and subjects.**

Two-Factor Within-*S* Design. We will now see how the rule generalizes to factorial designs. Let us first look at a factorial design with two within-*S* factors. Space does not permit a formal discussion of the analysis of this particular design, but examples of the design are often found in the research literature. This design is a two-way factorial, symbolized as an $(A \times B \times S)$ design, in which all subjects receive all the $(a)(b)$ treatment conditions. The order of the treatments is usually randomized or varied in some systematic fashion designed to minimize undesired sequence effects.

[5] If heterogeneity of within-group variances is present, it may be necessary to use specific error terms to evaluate treatment effects based on *portions* of the data, e.g., simple effects and interaction comparisons. In these cases, a useful procedure is to base the error term on only those observations involved in the calculation of the effect under consideration.

The standard sources of variability examined in the overall analysis of this (or any) two factor design are the two main effects (*A* and *B*) and the *A* × *B* interaction. In the $(A \times B \times S)$ design, these treatment sources are based on different configurations of the within-*S* factors. For example, the levels of factor *B* are totally disregarded in the calculation and assessment of the *A* main effect; for analysis purposes, the configuration of the data is equivalent to a single-factor within-*S* design involving the manipulation of factor *A*. Similarly, the levels of factor *A* are totally disregarded in the calculation and assessment of the *B* main effect; the configuration of the data in this case is equivalent to another single-factor within-*S* design involving the manipulation of factor *B*. Finally, *both* within-*S* factors are involved in the calculation and assessment of the *A* × *B* interaction. Applying the rule to each of these sources produces three different error terms:

> The error term for evaluating the *A* main effect is based on the interaction between factor *A* and subjects—the *A* × *S* interaction.
>
> The error term for evaluating the *B* main effect is based on the interaction between factor *B* and subjects—the *B* × *S* interaction.
>
> The error term for evaluating the *A* × *B* interaction is based on the interaction between factors *A* and *B* and subjects—the *A* × *B* × *S* interaction.

In each case, then, the error term consists of the interaction of the within-*S* factor (or factors) with subjects. The overall analysis is summarized in Table 19-5.

Three or More Within-*S* Factors. We are now in a position to describe the analysis of "pure" within-*S* factorials with any number of independent variables. For the overall analysis, all we need to do is to apply the general rule for within-*S* error

Table 19-5
Standard Analysis of the $(A \times B \times S)$ Design

Source	*df*	Error Term
A	$a - 1$	$A \times S$
B	$b - 1$	$B \times S$
S	$s - 1$	
$A \times B$	$(a - 1)(b - 1)$	$A \times B \times S$
$A \times S$	$(a - 1)(s - 1)$	
$B \times S$	$(b - 1)(s - 1)$	
$A \times B \times S$	$(a - 1)(b - 1)(s - 1)$	

terms to the specific design under consideration. That rule, to repeat, states that

> **The error term for evaluating any source of variability is based on an interaction between the factor or factors contained in the source and subjects.**

To illustrate with the three-factor design,

The error terms for the main effects of A, B, and C are $A \times S$, $B \times S$, and $C \times S$, respectively.
The error terms for the two-way interactions of $A \times B$, $A \times C$, and $B \times C$ are $A \times B \times S$, $A \times C \times S$, and $B \times C \times S$, respectively.
The error term for the $A \times B \times C$ interaction is $A \times B \times C \times S$.

The Mixed Two-Factor Design: A Review

The analysis of mixed factorial designs consists of a blending of a "pure" between-S design and a "pure" within-S design. We discussed this blending of designs in Chap. 17, when we considered the analysis of the mixed two-factor design in which A is the between-S factor and B is the within-S factor. (See Table 17-2, page 298, for a summary.) We will review this analysis here to illustrate how the general principles we have established for "pure" between-S and within-S designs apply to the analysis of the mixed factorial design.

A convenient way to understand the analysis of a mixed factorial is to segregate the treatment sources of variability into two categories, one representing the between-S portion of the analysis and the other the within-S portion. The normal "yield" of treatment effects for a two-factor design—A, B, and $A \times B$—and their error terms are listed in the first and second columns of Table 19-6, respectively.

Table 19-6
Error Terms for the $A \times (B \times S)$ Mixed Factorial Design

Treatment Source	Error Term
Between-Subjects Factor	
A	S/A
Within-Subjects Factor	
B	$B \times S/A$
$A \times B$	$B \times S/A$

You should note that the treatment sources of variability have been segregated into two categories. The first category consists of treatment sources based only on the between-S factor. In this particular design only one treatment source qualifies for membership in this category, namely, the A main effect. The error term for this between-S source is the same error term that would be appropriate if the data were collapsed over the within-S factor, in which case the arrangement would be a completely randomized design—a "pure" between-S design without repeated measures. This operation would produce a single-factor experiment in which S/A, the within-groups source, is the error term.

The second category consists of treatment sources based entirely or in part on the within-S factor in this design (factor B). There are two treatment sources that involve factor B: the B main effect and the $A \times B$ interaction. The error term for either of these within-S sources is essentially the same term that would be appropriate in a corresponding within-S design involving factor B. In the case of a "pure" $(B \times S)$ design, the error term would be the $B \times S$ interaction. In the mixed factorial, there is a different $B \times S$ interaction for each of the independent groups—a $B \times S$ interaction at level a_1 $(B \times S/A_1)$, a $B \times S$ interaction at level a_2 $(B \times S/A_2)$, and so on. These are combined and averaged to form the $MS_{B \times S/A}$ that serves as the error term for both B and $A \times B$ in the mixed factorial design.

Mixed Three-Factor Designs

We will now apply this system to the analysis of mixed three-factor designs. There are two types of mixed designs logically possible with three-factor designs. One of these contains two between-S factors and one within-S factor. If we designate factors A and B as the between-S factors and factor C as the within-S factor, we can refer to the arrangement as an $A \times B \times (C \times S)$ design, the parentheses indicating the portion of the design represented by repeated measures. The other type of mixed design contains one between-S factor and two within-S factors. If we designate factor A as the between-S factor and factors B and C as the within-S factors, we can refer to the arrangement as an $A \times (B \times C \times S)$ design, the parentheses again indicating the portion of the design represented by repeated measures.

Two Between-S Factors and One Within-S Factor. We can create an example of the first type of design by bringing together three of the independent variables we have introduced previously in our fictitious vocabulary experiment. Consider an experiment in which factor A consists of three different lectures (physical science, social science, and history) and factor B consists of two methods of presentation (computer and standard). Subjects are randomly assigned to independent groups