TWO-WAY MIXED DESIGN ANOVA

You will need to use the following from previous chapters:

**Symbols**

- \( k \): Number of independent groups in a one-way ANOVA
- \( c \): Number of levels (i.e., conditions) of an RM factor
- \( n \): Number of subjects in each cell of a factorial ANOVA
- \( N_i \): Total number of observations in an experiment
- \( e \): Coefficient to estimate the degree of sphericity in the population

**Formulas**

- Formula 14.2: \( SS_{	ext{error}} \) (by subtraction)
- Formula 14.3: \( SS_{	ext{total}} \) or one of its components

**Concepts**

Advantages and disadvantages of the RM ANOVA
- \( SS \) components of the one-way RM ANOVA
- \( SS \) components of the two-way ANOVA
- Interaction of factors in a two-way ANOVA

If you want the economy of a two-way factorial design, or its ability to detect the interaction of two independent variables, and at the same time you want the added power of repeated measures, you can use a two-way repeated measures design. The analysis of this design, using a two-way RM ANOVA, parallels the analysis of a three-way factorial design, just as the one-way RM ANOVA resembles the two-way ANOVA for independent groups. However, at least as common as the two-way RM design is a two-way factorial design in which one of the factors involves repeated measures (or matched subjects) and the other factor involves independent groups of subjects. For obvious reasons, this design is often called a mixed design, although this designation is not universal. Mixed designs are sometimes called split-plot designs, a description that arises from their early use in agricultural research. A potential source of confusion is that the term mixed design ANOVA sounds similar to the term mixed model ANOVA, which denotes a mixing of random and fixed effects in the same design. However, random effects are rarely used in psychological research (except for “subjects,” but that factor is almost never tested for significance), so there are few occasions for confusion. In this text, all of the ANOVA effects tested in earlier chapters were fixed effects, and that will continue to be the case for the remainder of the text.

Because mixed designs can have any number of within-subjects factors and between-subjects factors, it is usually necessary to specify the number of each type of factor. For instance, a three-way mixed design can have either two RM factors and one between-groups factor or one RM factor and two between-groups factors. In this chapter you will encounter only the simplest mixed design: the two-way mixed design. So long as the appropriate assumptions can be made, the data from a mixed design experiment can be analyzed with the mixed design ANOVA described in this chapter.

**The One-Way RM ANOVA Revisited**

To demonstrate how the mixed design ANOVA works, I will return to the experiment described at the beginning of Chapter 15, in which each of six
subjects must try to recall three types of words: positive, negative, and neutral. I will change the data, however, to suit the purposes of this chapter; see Table 16.1.

<table>
<thead>
<tr>
<th>Subject No.</th>
<th>Neutral</th>
<th>Positive</th>
<th>Negative</th>
<th>Row Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>21</td>
<td>17</td>
<td>19.33</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>18</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>15</td>
<td>18</td>
<td>16.67</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>10</td>
<td>13</td>
<td>12.67</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>4</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>7</td>
<td>4</td>
<td>6.33</td>
</tr>
<tr>
<td>Column Means</td>
<td>14.33</td>
<td>12.5</td>
<td>12.17</td>
<td>13</td>
</tr>
</tbody>
</table>

Table 16.1

First I will conduct a one-way repeated measures ANOVA for these data. As usual I will begin by finding $SS_{total}$. I can use Formula 14.3, and find that $SS_{total} = N\sigma^2(scores) = 18(25.89) = 466$. Alternatively, I can obtain the same answer with Formula 3.11:

$$SS_{total} = \sum X^2 - N\bar{X}_D^2 = 3,508 - 18(13^2) = 3,508 - 3,042 = 466$$

Next I will find $SS_{RM}$, which involves the means for each treatment (the column means in Table 16.1):

$$SS_{RM} = N\sigma^2(column means) = 18 \cdot \sigma^2(14.33, 12.5, 12.17) = 18(903) = 16.25$$

(Remember that you can use Formula 3.13A or an equivalent formula to calculate the biased variance of the column means if you do not have a statistical or scientific calculator.) Now I need to calculate $SS_{sub}$ so it can be subtracted from the total. I will use the means for each subject (the row means in Table 16.1):

$$SS_{sub} = N\sigma^2(row means) = 18 \cdot \sigma^2(19.33, 15, 16.67, 12.67, 8, 6.33) = 18(21.19) = 381$$

Finally, I can find $SS_{residual}$ (i.e., $SS_{inter}$) by subtraction:

$$SS_{residual} = 466 - 16.25 - 381.4 = 68.35$$

To find the $F$ ratio, first I find that $MS_{RM} = SS_{RM}/df_{RM} = 16.25/2 = 8.13$, and $MS_{residual} = SS_{residual}/df_{residual} = 68.35/10 = 6.84$. Therefore, $F = 8.13/6.84 = 1.19$. The results of the RM ANOVA are summarized in Table 16.2. Because the critical $F_M(2, 10)$ equals 4.1, we cannot reject the null hypothesis in this case.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between-Subjects</td>
<td>381.4</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within-Subjects</td>
<td>84.6</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>16.25</td>
<td>2</td>
<td>8.13</td>
<td>1.19</td>
</tr>
<tr>
<td>Residual</td>
<td>68.35</td>
<td>10</td>
<td>6.84</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>466</td>
<td>17</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
As you may have guessed, this is not the end of the story for this experiment. What I didn't tell you is that three of the subjects were selected because of high scores on a depression inventory, whereas the remaining three showed no signs of depression. The depressed subjects are graphed as dashed lines in Figure 16.1; solid lines represent the nondepressed subjects.

Converting the One-Way RM ANOVA to a Mixed Design ANOVA

The one-way RM ANOVA just performed can be transformed into a two-way mixed ANOVA by adding a between-groups factor with two levels: depressed and nondepressed. However, the advantage of this design may not be obvious from looking at Figure 16.1. On average, the two groups do not differ much in overall recall ability, nor does the variability within each group seem less than the total variability. The difference between the groups becomes apparent, though, when you focus on the subject × treatment interaction.

In Figure 16.2, the two groups are graphed in separate panels to make it obvious that the subject × treatment (S × T) interaction within each group is much smaller than the total amount of interaction when all subjects are considered together. The calculation of the mixed ANOVA can take advantage of this smaller S × T interaction by analyzing the SS components of the RM ANOVA further.

Analyzing the Between-Subjects Variability

Even though we would normally know about the depressed and nondepressed subgroups before analyzing our data, it would be reasonable to begin the mixed ANOVA by ignoring this distinction and calculating the RM ANOVA as we just did. The next step is to analyze further the between-subjects SS component shown in Table 16.2. This is the component we
As you may have guessed, this is not the end of the story for this experiment. What I didn't tell you is that three of the subjects were selected because of high scores on a depression inventory, whereas the remaining three showed no signs of depression. The depressed subjects are graphed as dashed lines in Figure 16.1; solid lines represent the nondepressed subjects.

**Converting the One-Way RM ANOVA to a Mixed Design ANOVA**

The one-way RM ANOVA just performed can be transformed into a two-way mixed ANOVA by adding a between-groups factor with two levels: depressed and nondepressed. However, the advantage of this design may not be obvious from looking at Figure 16.1. On average, the two groups do not differ much in overall recall ability, nor does the variability within each group seem less than the total variability. The difference between the groups becomes apparent, though, when you focus on the subject × treatment interaction.

In Figure 16.2, the two groups are graphed in separate panels to make it obvious that the subject × treatment (S × T) interaction within each group is much smaller than the total amount of interaction when all subjects are considered together. The calculation of the mixed ANOVA can take advantage of this smaller S × T interaction by analyzing the SS components of the RM ANOVA further.

**Analyzing the Between-Subjects Variability**

Even though we would normally know about the depressed and nondepressed subgroups before analyzing our data, it would be reasonable to begin the mixed ANOVA by ignoring this distinction and calculating the RM ANOVA as we just did. The next step is to analyze further the between-subjects SS component shown in Table 16.2. This is the component we...
This $F$ value (.059) allows us to test whether the means for the two groups differ significantly.

Because we are conducting a factorial ANOVA, we can say that this $F$ ratio tests the main effect of depression. Clearly, we do not have to look up a critical $F$ to know that the null hypothesis cannot be rejected. The fact that our observed $F$ is much less than 1.0 (indeed, unusually so) is of no interest to us. This just tells us that the means for the depressed and nondepressed subjects are surprisingly close together, given all of the variability within each group.

**Analyzing the Within-Subjects Variability**

Although we already calculated the $F$ for the main effect of word type when we performed the one-way RM ANOVA, this $F$ ratio must be recalculated for the mixed design to take into account the separation of subjects into subgroups. As you might have guessed, the numerator of the $F$ ratio for word type won’t change; this numerator depends on the separation of the means for the three word types, and it does not change just because we have regrouped the subjects within conditions. On the other hand, the denominator (the error term) does change, for reasons that can be seen by comparing Figure 16.2 to Figure 16.1. Notice that the subject $\times$ treatment interaction is fairly small within each group (Figure 16.2), but it looks rather large when all subjects are considered together (Figure 16.1). We can say that most of the $S \times T$ interaction is really due to a group by word type interaction, which should be removed from the total $S \times T$ interaction. This is the same as taking the two much smaller interactions of the subgroups and averaging them. Mathematically, what we need to do is further analyze the $SS_{\text{residual}}$ from the RM ANOVA into smaller components.

To subtract the group by word type interaction from the subject by word type interaction, we must calculate the former. This is done as it is for any two-way ANOVA and involves calculating the $SS_{\text{between-cells}}$ component. For this step we need to know the mean for each combination of group and word type. Table 16.3 shows the data from Table 16.1 rearranged into cells, with the cell means.

<table>
<thead>
<tr>
<th>Group</th>
<th>Neutral</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nondepressed</td>
<td>20</td>
<td>21</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>18</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Cell Means</td>
<td>14.67</td>
<td>15.33</td>
<td>10.67</td>
</tr>
<tr>
<td>Depressed</td>
<td>17</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Cell Means</td>
<td>14</td>
<td>9.67</td>
<td>13.67</td>
</tr>
</tbody>
</table>

The cell means from this table can then be inserted into Formula 14.3, as follows:

$$SS_{\text{between-cells}} = N\sigma^2(\text{cell means})$$

$$= 18 \cdot \sigma^2(14.67, 15.33, 10.67, 14, 9.67, 13.67)$$

$$= 18(4.364) = 78.55$$
We have already calculated the SS for word type (i.e., SS_{VT}) and the SS for groups, so we are ready to find the SS for the group by treatment interaction (SS_{G \times RM}) by subtraction.

\[ SS_{G \times RM} = 78.55 - 16.25 - 5.54 = 56.76 \]

Now we can go back to the original SS_{reid} from the one-way RM ANOVA and subtract SS_{G \times RM} to get the new smaller SS_{reid} for the mixed design (the sum of the subject by treatment interactions within each group). This new error component can be referred to simply as SS_{reid} or, more specifically, as the SS for the interaction of the subject factor with the repeated-measures factor, SS_{S \times RM}.

\[ SS_{S \times RM} = 68.35 - 56.76 = 11.59 \]

Now we have calculated all the SS components needed to complete the mixed design analysis. To summarize: The one-way RM ANOVA gave us three components, SS_{RM}, SS_{sub}, and SS_{reid}. SS_{sub} was then further divided to yield SS_{groups} and SS_{w}. SS_{reid} was also divided into two components: SS_{G \times RM} and SS_{S \times RM}. SS_{RM} was left alone. We have already tested the main effect of groups by forming an F ratio from MS_{groups} and MS_{w}. Now we can recalculate the F ratio for the repeated factor. The numerator MS is the same as in the one-way analysis: 8.13. But the new error term is based on SS_{S \times RM} (11.59) divided by df_{S \times RM} (8). So MS_{S \times RM} = 11.59/8 = 1.45. Thus the new F for testing the main effect of word type is

\[ F = \frac{MS_{S \times RM}}{MS_{S \times RM}} = \frac{8.13}{1.45} = 5.61 \]

The critical F is based on two and eight degrees of freedom; F_{.05} = 4.46. Because of the smaller error term in the mixed design, the main effect of word type is now significant at the .05 level.

**Two-Way Interaction in the Mixed Design ANOVA**

Like any other two-way ANOVA, the mixed design ANOVA can give us one more F ratio: a test of the interaction of the two factors. We find MS_{G \times RM} by dividing SS_{G \times RM} by df_{G \times RM}. We have already found that SS_{G \times RM} = 56.76, and you will have to take on faith for the moment that df_{G \times RM} = 2. So MS_{G \times RM} = 56.76/2 = 28.38. (Formulas for the mixed design ANOVA, including those for the df, will be presented more systematically in Section B.) To complete the F ratio, however, we need to know which MS to use as the error term. So far both MS_{w} and MS_{S \times RM} have been used as error terms. Can we use one of these error terms, or is there some third error term to use? The answer is that MS_{S \times RM} is the appropriate error term, for reasons that I will make clear shortly. To test whether depression and word type interact significantly, we form the following F ratio:

\[ F = \frac{MS_{G \times RM}}{MS_{S \times RM}} = \frac{28.38}{1.45} = 19.57 \]

You should not have to look up a critical F to know that such a large observed F ratio must be significant at the .05 level (except in the ridiculous case when some of your groups contain only one subject).

To understand why MS_{S \times RM} is the appropriate error term for the interaction of the two factors, it may help to compare a graph of the cell means...
(see Figure 16.3) to a graph of the data from individual subjects (see Figure 16.2). Notice that each line in Figure 16.3 represents the average number of words recalled by the subjects in that group. To the extent that the individual subjects in a group are not parallel to the group average, or from each other, $MS_{S \times RM}$ increases. As $MS_{S \times RM}$ increases, so does the chance of finding an accidental interaction between the groups (just as increased variation within groups leads to a greater chance of finding a large difference among group means in an independent ANOVA). Thus, $MS_{G \times RM}$ contains error arising from $MS_{S \times RM}$ in addition to any real interaction between the factors. When the null hypothesis is true, there is no real interaction contributing to $MS_{G \times RM}$, so the $F$ ratio—$MS_{G \times RM}/MS_{S \times RM}$—should equal about 1.0.

**Summarizing the Mixed Design ANOVA**

The results of our mixed ANOVA are summarized in Table 16.4. The structure of the table tells us a great deal about the structure of our analysis. First, the table is broken into two distinct sections. The upper section, "Between-Subjects," deals with the variation of scores within each column of Table 16.1; if all the scores were the same for a particular word type, the $SS$ components for this section would all be zero. The differences that do exist between subjects can be divided into the difference between the subgroups ($SS_{groups}$) and the differences between subjects within each group ($SS_{wb}$); the latter forms the basis of the error term for this part of the table. The number of degrees of freedom for between-subjects variation is one less than the total number of different subjects, which equals $6 - 1 = 5$. These five df break down into one df for group differences ($k - 1 = 2 - 1 = 1$) and four df for differences within groups (number of subjects − number of groups = $6 - 2 = 4$).

The lower section of Table 16.4, "Within-Subjects," deals with the variation of scores within each row of Table 16.1; if each subject produced the same recall score for all word types, the $SS$ components in this section of the table would all be zero (even though subjects might differ from each
The differences that do exist between levels of the repeated factor (in this case, word type) can be divided into average differences among the word types (SSRM), the interaction of groups with word type (SSg×RM), and the interaction of individual subjects with word type (SSg×RM). For instance, if a particular subject recalls more positive than negative words, it may be because subjects tend to do this in general or because subjects in a particular subgroup tend to do this or because this particular subject has an individual tendency to do this. The last of these sums of squares (SSg×RM) forms the basis of the error term for the lower section of Table 16.4. The number of degrees of freedom for within-subjects variation equals the total number of subjects (six) times one less than the number of repeated measures (6 – 1 = 5), which equals 12. These 12 df break down into 2 df for treatment differences (c – 1), 2 df for the group by treatment interaction [(k – 1)(c – 1)], and the remainder, 8 df, for the residual, which is the number of groups times one less than the number of repeated conditions times one less than the number of subjects in each subgroup.

**Interpreting the Results**

The between-groups factor in our example was not found to be significant, whereas the within-subjects factor was. This pattern is probably more common than the reverse pattern because the test of the repeated factor is likely to have greater power, for reasons discussed in Chapter 15. However, we should interpret the results of the main effects cautiously because of the significance of the interaction. The effect of the repeated factor is different for the two subgroups, and further analysis would be appropriate to localize these differences. As with other two-way ANOVAs, a graph of the cell means can help you understand how the main effects combine with the interaction to produce the pattern of results. You can see from Figure 16.3 that the depressed subjects have relatively poor recall for positive words, whereas it is the recall of negative words that is weak for nondepressed subjects. These particular comparisons can be tested by methods I will discuss in Section C.

**The Varieties of Mixed Designs**

**The Between-Groups Factor Is a Grouping Variable**

In the example above, the RM factor involves an experimental manipulation (i.e., type of word), whereas the between-groups factor is based on preexisting individual differences in depression. This is a common form of the mixed design. The reduction in SSRM caused by separating the depressed and nondepressed subjects is reminiscent of the reduction in SSRW that followed the separation of men and women in the example of a depression
Chapter 16 • Two-Way Mixed Design ANOVA

drug in Chapter 14. This reduction will occur whenever there is an interaction between the experimental and grouping factor. However, adding a grouping variable causes a reduction in df for the error term (from 10 to 8 in the example above), which means that \( S_{e} \) is being divided by a smaller number, and that tends to increase your error term. Therefore, if the interaction is quite small, adding the grouping variable can actually lower your \( F \) ratio. The lesson is that grouping variables should not be added casually.

Often, the reason a researcher adds a grouping variable is that he or she expects the between-groups factor to interact with the RM factor; in such cases, there is little interest in the main effect of the between-groups factor. The chief purpose of such a design is to determine whether the effects of the RM factor are the same for various subgroups in the population. In the preceding example, it is of interest that the types of words remembered most easily depends on the subjects’ level of depression.

**The Between-Groups Factor Is an Experimental Variable**

Another type of mixed design arises when the between-groups factor involves an experimental manipulation that does not lend itself to repeated measures. For instance, consider an experiment in which each subject completes a series of tasks of varying difficulty. In one condition, the subjects are told that the tasks come from a test of intelligence and that their performance will give an indication of their IQ. In another condition, subjects are given monetary rewards for good performance, and in a third condition, subjects are simply asked to work as hard as they can. It should be clear that once the researcher has run a subject in the IQ condition, running the same subject again in a different condition (using similar tasks) would yield misleading results. Nor could you always run the IQ condition last without confounding your results with an order effect. A reasonable solution is the mixed design, with the three motivational conditions as the levels of a between-groups factor, and task difficulty as a within-subjects factor.

One purpose of the preceding experiment could be to explore the interaction between the motivational condition and task difficulty. (For instance, are performance differences between difficulty levels the same for different types of motivation?) However, if there were no interaction, it might still be interesting to examine the main effect of motivational condition. (The main effect of difficulty would merely confirm that the difficulty manipulation worked.) In general, when there is no interaction between the factors of a mixed design, the analysis reduces to two one-way ANOVAs: an independent-groups ANOVA and an RM ANOVA.

In one common form of the mixed design the repeated-measures factor is merely the passage of time; this type of RM factor is rarely used unless the between-groups factor involves an experimental manipulation (e.g., measurements can be taken before, during, and/or after some treatment, with different groups assigned to different treatments). An example of this type of design will be used in Section B to illustrate the systematic calculation of a two-way mixed design.

**The RM Factor Can Be Based on Repeated Measures or Randomized Blocks**

The levels of the RM factor in a mixed design can be administered in several ways, as discussed for the one-way RM design. The levels can be interspersed for a presentation that is virtually simultaneous, or the levels can
be presented successively. Successive presentations usually produce order effects that can be averaged out by counterbalancing but nonetheless inflate the error term and decrease power. Special ANOVA techniques can be employed to remove the influence of order effects from the error term, as I will describe in Section C. Those special techniques can also be used to quantify the magnitude of differential carryover effects. If carryover effects are severe and unavoidable, the experiment should be designed to match subjects into blocks; randomized blocks can provide much of the benefit of repeated measures with no order effects at all. Finally, if carryover cannot be avoided, and no basis for matching can be found, both factors would have to consist of independent samples, and the two-way ANOVA described in Chapter 14 would be appropriate.

1. The two-way mixed design (also called a split-plot design) includes one between-subjects factor and one within-subjects factor; the latter involves either repeated measures or matched blocks of subjects. The total variability in a two-way mixed design can be initially divided into between-subjects and within-subjects variation.

2. The between-subjects variation can be further subdivided into a portion that depends on the separation of the group means and a portion that depends on subject-to-subject variation within each group. The within-group variability is used as the error term for testing the main effect of the between-subjects factor.

3. The within-subjects variation can be divided into three components: one that reflects variation among the means of the repeated conditions, one that reflects the interaction of the two factors, and one that reflects the interaction of subjects with the RM factor within each group. The last of these components is used as the error term for testing both the main effect of the within-subjects factor and the interaction of the two factors.

4. Adding a between-subjects factor to a one-way RM design is likely to reduce the error term of the RM factor if the between-subjects factor is based on preexisting groups and there is some interaction between the two factors. Adding the grouping factor also allows you to test the interaction of the two factors and thereby determine whether the effect of the RM factor is similar for different subgroups of the population. However, adding the grouping factor reduces the degrees of freedom for the error term, so it can be counterproductive if the grouping variable is not related to your dependent variable.

5. A mixed design commonly arises by adding the passage of time (RM factor) to a between-subjects factor that is based on an experimental manipulation. In such a design, it is usually only the interaction of the two factors that is interesting. Another common type of mixed design is one in which one factor lends itself easily to repeated measures (trials at different levels of difficulty, randomly interspersed), but the other factor does not (telling some subjects that the task tests intelligence, and others that it does not).

6. The within-subjects (RM) factor in a mixed design has the same advantages and disadvantages as the within-subjects factor in a one-way RM design. This factor is likely to have greater power than the between-subjects factor because subject-to-subject variability is ignored; however, if the repeated measures are not simultaneous, you may need to counterbalance, and if differential carryover effects are likely, you may need to match subjects in blocks as an alternative to repeated measures.
1. a. Devise a mixed design experiment in which the between-subject variable is quasi-independent.
    b. Devise a mixed design experiment in which the between-subjects variable is manipulated by the experimenter.
    c. Devise a mixed design experiment in which the within-subjects variable involves matched subjects rather than repeated measures.

2. A researcher tested two groups of subjects—six alcohol abusers and six moderate social drinkers—on a reaction time task. Each subject was measured twice: before and after drinking 4 ounces of vodka. A mixed-design ANOVA produced the following SS components: SS_{groups} = 88, SS_{w} = 1380, SS_{RM} = 550, SS_{G \times RM} = 2.0, and SS_{S \times RM} = 134. Complete the analysis and present the results in a summary table.

3. Exercise 15B4 described a randomized-blocks experiment involving four different textbooks and nine blocks of subjects. The RM ANOVA produced the following SS components: SS_{within} (SS_{RM}) = 76.75, SS_{subject} = 612.5, and SS_{residual} = 27.5. Now suppose that the nine blocks of subjects can be separated into three subgroups on the basis of overall ability and that the mixed design ANOVA yields SS_{groups} = 450 and SS_{G \times RM} = 8.5. Complete the analysis and present the results in a summary table.

4. The following table shows the number of ounces of popcorn consumed by each subject while viewing two emotion-evoking films, one evoking happiness and one evoking fear.

<table>
<thead>
<tr>
<th></th>
<th>Happiness</th>
<th>Fear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preload</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>No</td>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td>Load</td>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>15</td>
</tr>
</tbody>
</table>

   a. Does there appear to be about the same amount of subject \times treatment interaction in each group? 
   b. Does there appear to be a considerable amount of group \times repeated-measure interaction?

5. If you calculate an RM ANOVA and then assign the subjects to subgroups to create a mixed design, the observed $F$ ratio for the RM factor may get considerably larger. Under which of the following conditions is this likely?
   a. The degrees of freedom associated with the error term are reduced considerably.
   b. There is a good deal of subject \times RM treatment interaction.
   c. There is a good deal of (sub)group \times RM treatment interaction.
   d. There is a good deal of subject to subject variability.

In Chapter 11, I pointed out the weakness of the simple before-after design. Even if the before-after difference turns out to be statistically significant for some treatment, without a control group it is difficult to specify the cause of the difference. Was the treatment really necessary to produce the difference, or would just the act of participating in an experiment be sufficient? When you add a control group and continue to measure your variable twice in both groups, you have created a mixed design. For example, suppose that you have devised a new treatment for people afraid of public speaking. To show that the effects of your new treatment are greater than a placebo effect, half the subjects (all of whom have this phobia) are randomly assigned to a control group; their treatment consists of hearing inspirational talks about the joys of public speaking. Suppose further that you wish to demonstrate that the beneficial effects of your treatment last beyond the end of the treatment period. Consequently, you measure the
degree of each subject’s phobia not only before and after the treatment period, but also 6 months after the end of treatment (follow-up). We will apply our usual six-step hypothesis testing procedure to this experiment.

Step 1. State the Hypotheses

The design in this example consists of two factors that have been completely crossed (i.e., a two-way factorial design). As such, the design involves three independent null hypotheses that can be tested. The \( H_0 \) for the main effect of treatment (the between-subjects factor) is \( \mu_{\text{exp}} = \mu_{\text{con}} \). The \( H_0 \) for the main effect of time (the within-subjects, or repeated, factor) is \( \mu_{\text{pre}} = \mu_{\text{post}} = \mu_{\text{final}} \). The \( H_0 \) for the interaction of the two factors is a statement that the experimental-control difference will be the same at each point in time (before, after, and at follow-up) or, more simply, that the effects of one factor are independent of the effects of the other factor. The alternative hypothesis for the main effect of treatment can be stated either as one-tailed (e.g., \( \mu_{\text{exp}} > \mu_{\text{con}} \)) or as two-tailed (\( \mu_{\text{exp}} \neq \mu_{\text{con}} \)). I will take the more conservative approach and use the two-tailed \( H_a \). For the main effect of time there are three levels, so the only simple way to state \( H_a \) is to state that \( H_a \) is not true. Also, for simplicity the \( H_a \) for the interaction is a statement that the corresponding \( H_0 \) is not true.

Step 2. Select the Statistical Test and the Significance Level

The time factor involves three measures on each subject, but the treatment factor involves different subjects. Because our purpose is to detect differences in population means along these two different dimensions, a mixed design ANOVA is appropriate. The conventional approach is to use .05 as alpha for all three null hypothesis tests.

Step 3. Select the Samples and Collect the Data

To minimize the calculations I will assume that only eight phobic subjects are available for the experiment and that four are selected at random for each treatment group. The dependent variable will be a 10-point rating scale of phobic intensity with respect to public speaking (from 0 = relaxed when speaking in front of a large audience to 10 = incapable of making a speech in front of more than one person). Because each subject is measured three times, there will be a total of \( 8 \times 3 = 24 \) ratings or observations, as shown in Table 16.5.

<table>
<thead>
<tr>
<th>Group</th>
<th>Subject No.</th>
<th>Before</th>
<th>After</th>
<th>Follow-up</th>
<th>Row Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phobia</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Treatment</td>
<td>2</td>
<td>9</td>
<td>6</td>
<td>5</td>
<td>6.67</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>5</td>
<td>4.67</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>5.33</td>
</tr>
<tr>
<td>Control</td>
<td>5</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Treatment</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>7.33</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>6.67</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>7</td>
<td>5.67</td>
</tr>
<tr>
<td>Cell Means</td>
<td>7.25</td>
<td>6.25</td>
<td>7.25</td>
<td></td>
<td>6.92</td>
</tr>
<tr>
<td>Column Means</td>
<td>7.375</td>
<td>5.375</td>
<td>6.125</td>
<td></td>
<td>6.292</td>
</tr>
</tbody>
</table>
Chapter 16 • Two-Way Mixed Design ANOVA

Step 4. Find the Regions of Rejection

Given that certain assumptions have been met (these will be discussed shortly), it is appropriate to use the $F$ distribution to find a critical value for each null hypothesis. However, we need to know the degrees of freedom that apply in each case. The breakdown of the df can get complicated for a mixed design, so a df tree can be especially helpful when dealing with this type of design (see Figure 16.4). In Figure 16.4, I use $k$ to represent the number of different groups (i.e., the number of levels for the between-groups factor), $c$ to represent the number of treatment conditions presented to each subject (i.e., the number of levels for the within-subjects factor), and $n$ to represent the total number of cells in the two-way mixed design. I will deal only with the case in which each group has the same number of subjects, so $n$ can be used to represent the number of subjects assigned to each level of the between-groups factor (i.e., the number of subjects per group); thus $kn$ is the total number of different subjects in the experiment (I called this $N$ without a subscript, in Section A). However, there are $c$ measurements, or scores, for each subject, so $nkc$ is the total number of observations in the experiment (i.e., $N_T$). Now we can express the total df as

$$\text{df}_{\text{tot}} = nkc - 1 \quad \text{or} \quad (N_T - 1)$$

The total df are divided into the df associated with variation between subjects ($\text{df}_{\text{between}}$) and the df associated with variation within subjects ($\text{df}_{\text{within}}$). The $\text{df}_{\text{between}}$ are simply the total number of different subjects minus 1; the $\text{df}_{\text{within}}$ are equal to the total number of observations minus the total number of different subjects. These relationships are expressed in the first two parts of Formula 16.1:

a. $\text{df}_{\text{between}} = nk - 1$

b. $\text{df}_{\text{within}} = nkc - 1$ or $nkc - nk$

As you can see from Figure 16.4, the df are further subdivided in each branch. The $\text{df}_{\text{between}}$ are divided into the following two components:

c. $\text{df}_{\text{groups}} = k - 1$

d. $\text{df}_{\text{W}} = k(n - 1)$ or $nk - k$

---

**Figure 16.4**

Degrees of Freedom Tree for Two-Way Mixed Design

[Diagram showing the breakdown of degrees of freedom with $nzc - 1$, $nk - 1$, $nkc - 1$, etc.]
Section B • Basic Statistical Procedures

where \( df_w \) stands for the df associated with subject-to-subject variation within each group. The \( df_{within-S} \) are divided into three components:

\[
e. \quad df_{RM} = c - 1
f. \quad df_{G \times RM} = (k - 1)(c - 1)
g. \quad df_{S \times RM} = k(c - 1)(n - 1)
\]

The df for the interaction of the two factors, \( df_{G \times RM} \), is equal to the df for groups times the df for the repeated measures factor. The \( df_{S \times RM} \) component is the df corresponding to the sum of the df for the subject by repeated measure interactions for each group.

For the present example, \( k = 2, c = 3 \), and \( n = 4 \), so \( df_f = 4 \cdot 2 \cdot 3 - 1 = 24 - 1 = 23 \), which can be broken down as follows:

\[
df_{between-S} = (4 \cdot 2) - 1 = 8 - 1 = 7
\]
\[
df_{within-S} = 4 \cdot 2(3 - 1) = 8 \cdot 2 = 16
\]

Then, \( df_{between-S} \) can be divided as follows:

\[
df_{groups} = 2 - 1 = 1
\]
\[
df_w = 2(4 - 1) = 6
\]

Similarly, \( df_{within-S} \) can be divided into the following components:

\[
df_{RM} = 3 - 1 = 2
\]
\[
df_{G \times RM} = 1 \cdot 2 = 2
\]
\[
df_{S \times RM} = 2(3 - 1)(4 - 1) = 2 \cdot 2 \cdot 3 = 12
\]

Now we can find the critical \( F \) value for each of our three null hypothesis tests. For the main effect of phobia treatment, the df are \( df_{groups} \) (1) and \( df_w \) (6); \( F_{adj}(1, 6) = 5.99 \). For the main effect of time, the df are \( df_{RM} \) (2) and \( df_{S \times RM} \) (12); \( F_{adj}(2, 12) = 3.89 \). And for the interaction of the two factors, the df are \( df_{G \times RM} \) (2) and \( df_{S \times RM} \) (12); therefore, the critical \( F \) for this test is also 3.89.

Step 5. Calculate the Test Statistics

For each of the df components delineated in the preceding, there is a corresponding SS component. As usual, these SS components will add up to \( SS_{total} \) which, for this example, equals \( N \cdot \sigma^2(\text{scores}) = 24(2.457) = 58.97 \). A convenient next step is to calculate \( SS_{between-S} \), which depends on the means for each subject (the eight row means in Table 16.5). I will be calculating all of the SS components with Formula 14.3, except for those which are more conveniently found by subtraction.

\[
SS_{between-S} = 24 \cdot \sigma^2(6, 6.67, 4.67, 5.33, 8, 7.33, 6.67, 5.67)
\]
\[
= 24(1.0386) = 24.93
\]

Now we find the SS for the grouping factor, which depends on the means for each group (these are the row means if you think of the rows as

\[
\frac{df_{G \times RM}}{k(c - 1)(n - 1)}
\]
consisting of cell means, rather than individual scores):

\[ SS_{\text{groups}} = N_f \sigma^2(\text{group means}) = 24 \cdot \sigma^2(5.67, 6.92) = 24(391) = 9.38 \]

By subtracting \( SS_{\text{groups}} \) from \( SS_{\text{between-S}} \) we obtain \( SS_w \) (Formula 16.2):

\[ SS_w = SS_{\text{between-S}} - SS_{\text{groups}} \quad \text{Formula 16.2} \]

For this example,

\[ SS_w = 24.93 - 9.38 = 15.55 \]

One branch of the total \( SS \) has now been analyzed into its components. To find the total of the other branch we subtract \( SS_{\text{between-S}} \) from \( SS_{\text{total}} \) to find \( SS_{\text{within-S}} \):

\[ SS_{\text{within-S}} = SS_{\text{total}} - SS_{\text{between-S}} \quad \text{Formula 16.3} \]

In this case,

\[ SS_{\text{within-S}} = 58.97 - 24.93 = 34.04 \]

We now turn our attention to analyzing \( SS_{\text{within-S}} \) into its components, starting with the \( SS \) for the repeated measures factor (\( SS_{\text{RM}} \)), which is based on the means of the columns in Table 16.5.

\[ SS_{\text{RM}} = N_f \sigma^2(\text{column means}) = 24 \cdot \sigma^2(7.375, 5.375, 6.125) \]
\[ = 24(6.806) = 16.3 \]

The \( SS \) for the interaction of the two factors (\( SS_{G \times RM} \)), as in any two-way ANOVA, requires that we calculate \( SS_{\text{within-cells}} \) and then subtract the \( SS \) components for the two main effects. Using the cell means from Table 16.5, we obtain:

\[ SS_{\text{within-cells}} = N_f \sigma^2(\text{cell means}) = 24 \cdot \sigma^2(7.5, 4.5, 5, 7.25, 6.25, 7.25) \]
\[ = 24(1.363) = 32.71 \]

Now we can use Formula 16.4 to obtain \( SS_{G \times RM} \):

\[ SS_{G \times RM} = SS_{\text{within-cells}} - SS_{\text{groups}} - SS_{\text{RM}} \quad \text{Formula 16.4} \]

For these data,

\[ SS_{G \times RM} = 32.71 - 9.38 - 16.33 = 7.0 \]

The third component of \( SS_{\text{within-S}} \) is also found by subtraction, according to Formula 16.5:

\[ SS_{S \times RM} = SS_{\text{within-S}} - SS_{\text{RM}} - SS_{G \times RM} \quad \text{Formula 16.5} \]

Therefore,

\[ SS_{S \times RM} = 34.04 - 16.33 - 7.0 = 10.71 \]

To obtain the variance estimates (i.e., \( MS \)) that we will need to form our three \( F \) ratios, we must divide each of the five final \( SS \) components we found above by their corresponding \( df \)’s as shown in Formula 16.6.
\[ MS_{\text{groups}} = \frac{SS_{\text{groups}}}{df_{\text{groups}}} \]
\[ MS_W = \frac{SS_W}{df_W} \]
\[ MS_{RM} = \frac{SS_{RM}}{df_{RM}} \]
\[ MS_{G \times RM} = \frac{SS_{G \times RM}}{df_{G \times RM}} \]
\[ MS_{S \times RM} = \frac{SS_{S \times RM}}{df_{S \times RM}} \]

Inserting the values for this example yields

\[ MS_{\text{groups}} = \frac{9.38}{1} = 9.38 \]
\[ MS_W = \frac{15.55}{6} = 2.59 \]
\[ MS_{RM} = \frac{16.33}{2} = 8.17 \]
\[ MS_{G \times RM} = \frac{7}{2} = 3.5 \]
\[ MS_{S \times RM} = \frac{10.71}{12} = .89 \]

Finally, we can form the \( F \) ratios to test each of the three null hypotheses stated in Step 1:

\[ F_{\text{groups}} = \frac{MS_{\text{groups}}}{MS_W} \]  
\[ F_{RM} = \frac{MS_{RM}}{MS_{S \times RM}} \]
\[ F_{G \times RM} = \frac{MS_{G \times RM}}{MS_{S \times RM}} \]

To test the main effect of groups, we create the following ratio:

\[ F_{\text{groups}} = \frac{9.38}{2.59} = 3.62 \]

To test the main effect of the within-subject factor, \( MS_{S \times RM} \) is the error term:

\[ F_{RM} = \frac{8.17}{.89} = 9.19 \]

The same error term is used in the \( F \) ratio to test the two-way interaction:

\[ F_{G \times RM} = \frac{3.5}{.89} = 3.93 \]

**Step 6. Make the Statistical Decisions**

The observed \( F \) for the main effect of the phobia treatment is 3.62, which is less than the critical \( F(5, 99) \), so we cannot reject the null hypothesis for this
factor. On the other hand, the $F$ ratio for the time factor (before versus after versus follow-up) is 9.19, which is well above the critical $F(3.89)$, so the null hypothesis for this effect can be rejected. For the interaction, the observed $F$ ratio (3.93) is only slightly above the critical $F(3.89)$, but that is all that is required to reject this null hypothesis, as well.

**Interpreting the Results**

At first, the lack of statistical significance for the main effect of treatment group may seem discouraging; it seems to imply that the phobia treatment didn’t work or that at best it was no more effective than the control procedure. However, the significant interaction should remind you to graph the cell means before trying to interpret the results of the main effects. You can see from Figure 16.5 that the two groups are very similar in phobic intensity before the treatment (which is to be expected with random assignment) but diverge considerably after treatment. Despite the similarity of the groups before treatment, the two later measurements might have caused the main effect of group to be significant had not the samples been so small. The $F$ ratio for groups is sensitive to the variability from subject to subject, and with small samples power is low unless the between-group effect size is quite large.

On the other hand, subject-to-subject variability does not affect the $F$ ratio for the time factor. As long as the subjects exhibit similar patterns over time within each group, $MS_{C \times RM}$ will tend to be small and $F_{RM}$ will tend to be large, as is the case in this example. However, the significance of the time factor must also be interpreted cautiously, given that the interaction is significant. The significance of $F_{G \times RM}$ tells you that the effect of time is different for the two groups. From Figure 16.5 you can see that the before-after reduction is much larger for the experimental group, and the increase from after to follow-up is somewhat greater for the control group.

It is likely that after obtaining these results a researcher would think of some more specific hypotheses to test—for instance, are the two groups significantly different just after the treatment? Or, is the before-after phobia

---

**Figure 16.5**

Graph of Cell Means for the Data in Table 16.5

<table>
<thead>
<tr>
<th>Phobic intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

- **Phobic treatment**
- **Control group**

Before | After | Follow-up
Section B - Basic Statistical Procedures

Publishing the Results of a Mixed ANOVA

The results of the preceding phobia treatment experiment could be reported in a journal article in this manner: "The phobia intensity ratings were submitted to a $2 \times 3$ mixed design ANOVA, in which treatment group (experimental versus placebo control) served as the between-subjects variable, and time (before versus after versus follow-up) served as the within-subjects variable. The main effect of treatment group did not attain significance, $F(1, 6) = 3.61$, $MSE = 2.6$, $p > .05$, but the main effect of time did reach significance, $F(2, 12) = 9.19$, $MSE = .89$, $p < .05$. The results of the main effects are qualified, however, by a significant group by time interaction, $F(2, 12) = 3.94$, $MSE = .89$, $p < .05$. The cell means reveal that the before-after decrease in phobic intensity was greater, as predicted, for the phobia treatment group and that this group difference was maintained at follow-up. In fact, at follow-up, the control group's phobic intensity had nearly returned to its level at the beginning of the experiment."

The preceding paragraph would very likely be accompanied by a table or graph of the cell means and followed by a report of more specific comparisons, such as testing for a significant group difference just at the follow-up point. Although it is not likely that an ANOVA summary table would be included in the report of your results, such tables are produced by most statistical software packages, and because they are instructive, I include a summary table (Table 16.6) for the phobia treatment example.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>$F$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between-Subjects</td>
<td>24.96</td>
<td>7</td>
<td>3.61</td>
<td>&gt;.05</td>
<td></td>
</tr>
<tr>
<td>Groups</td>
<td>9.38</td>
<td>1</td>
<td>9.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within-Groups</td>
<td>15.58</td>
<td>6</td>
<td>2.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within-Subjects</td>
<td>34</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>16.33</td>
<td>2</td>
<td>8.2</td>
<td>9.19</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Group $\times$ Time</td>
<td>7.0</td>
<td>2</td>
<td>3.5</td>
<td>3.94</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Residual ($S \times \text{Time}$)</td>
<td>10.67</td>
<td>12</td>
<td>.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>58.96</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Assumptions of the Mixed Design ANOVA

I didn't say anything about how phobia was measured in the preceding experiment, other than to say that a 10-point scale was used, but a basic assumption of all the parametric statistical tests in this text (i.e., all the tests covered thus far) is that the measurement scale used has the interval property, as defined in the first chapter. Nonetheless, as I have mentioned earlier, it is common in psychological research to perform parametric tests on data arising from subjective rating scales without demonstrating that the scale possesses the interval property. You will occasionally see the matter debated, but I will remain neutral on this matter. The other usual ANOVA assumptions apply as well to the mixed design; this includes independent
random sampling (but random assignment is usually substituted), normal distributions, and homogeneity of variance.

**Homogeneity of Variance**

When the design is balanced, there is usually little concern about this assumption. However, in a mixed design there are two very different ways in which the design can be unbalanced. One way involves the between-subjects factor. It is not uncommon to end up with different numbers of subjects in each independent group, either because you are dealing with intact groups (e.g., there may be different numbers of patients available for each diagnosis represented in the study) or because some subjects have to be deleted from the data after the study has been concluded. Because unequal group sizes complicate the analysis, especially if the within-group variances are quite different from one group to another, I will not be covering such cases in this chapter, but the analysis is fairly routine and similar to the unbalanced two-way independent ANOVA.

The other way the mixed design can become unbalanced involves one or more subjects missing data for one or more of the RM levels. This can occur when a subject drops out of an experiment before the final session or leaves out some responses on a questionnaire or provides data that is just not usable for some of the conditions. Missing RM data produces such statistical complications that most researchers either replace the subject entirely or fill in the missing values by using some technique to estimate the most likely value for each missing data point. Statistical methods for dealing with missing RM data are well beyond the scope of this text. When there is no missing data, the homogeneity of variance assumption with respect to the RM levels is not important if the sphericity assumption is satisfied, as described in the previous chapter.

**Homogeneity of Covariance Across Groups**

An assumption unique to mixed designs is that the covariance structure among the RM levels must be the same for all of the independent groups. In terms of the preceding example, the phobia group should have the same amount of before-after interaction (in the population) as the control group, and this should hold for the other two pairs of RM levels, as well. This form of homogeneity can be tested in terms of a statistic known as Box's M criterion (Huynh & Mandeville, 1979), which is provided automatically by some statistical packages (e.g., SPSS) whenever a mixed design ANOVA is requested.

**Sphericity in Mixed Designs**

If a test of Box’s M is not significant, the interactions for pairs of RM levels can be calculated by averaging across the different independent groups before testing for sphericity with Mauchly’s W statistic. If the W statistic is not significant either, one can proceed with the calculations of the mixed design ANOVA as demonstrated in this chapter. However, if W is significant or you are dealing with small sample sizes or nonnormal distributions, you should follow the procedures described in the next subsection. If Box’s M statistic is significant, the methods of this chapter are not justified, and you may have to proceed by performing separate one-way RM ANOVA’s for each of your independent groups, followed by cell-to-cell comparisons (with an appropriate adjustment of your alpha for each comparison).
Dealing with a Lack of Sphericity in Mixed Designs

As described in the previous chapter, a lack of sphericity increases the likelihood of a significant $F$ ratio, even if all of the population means for the repeated treatments are equal. A three-step procedure was described to keep the Type I error rate at the level set initially. This procedure can be applied to the RM factor of a mixed design, as well. If $F_{RM}$ does not meet the usual criterion for significance, you need do nothing further. For the conservatively adjusted critical $F$, $df_{num} = 1$ and $df_{denom} = df_{W} = k(n - 1)$. If the $F$ ratio for the RM factor ($F_{RM}$) surpasses this larger critical $F$, it can be considered statistically significant without assuming sphericity. However, if $F_{RM}$ falls between the usual critical $F$ and the conservative $F$, you need to compute the epsilon ($\epsilon$) coefficient to estimate the degree of sphericity in the population and adjust the degrees of freedom accordingly.

The $F$ ratio for the interaction of the two factors ($F_{G \times RM}$) uses the same error term as $F_{RM}$ and can therefore be similarly biased when sphericity is violated. The three-step procedure should also be applied to the test of $F_{G \times RM}$ except that the degrees of freedom for the numerator of the adjusted critical $F$ are $df_{num} = df_{groups} = k - 1$ ($df_{denom}$ still equals $df_{W}$). This modified univariate approach can be replaced, however, by a multivariate analysis of both the RM factor and the interaction of the two factors.

A Special Case: The Before-After Mixed Design

The simplest possible mixed design is one in which an experimental group and a control group are measured before and after some treatment. Some statisticians have argued that a two-way ANOVA is unnecessary in this case and may even prove misleading (Huck & McLean, 1975). The main effect of treatment is misleading because the group difference before the treatment (which is expected to be very close to zero) is being averaged with the group difference after the treatment. The main effect of time is equally misleading because the before-after difference for the experimental group is being averaged with the before-after difference for the control group. The only effect worth testing is the interaction, which tells us whether the before-after differences for the experimental group are different from the before-after differences for the control group. We do not need ANOVA to test this interaction—we need only find the before-after difference for each subject and then conduct a $t$ test of two independent samples (experimental versus control group) on these difference scores. Squaring this $t$ value will give the $F$ that would be calculated for the interaction in the mixed design ANOVA. If before and after measurements have been taken on more than two groups, a one-way independent-groups ANOVA on the difference scores yields the same $F$ ratio as a test of the interaction in the mixed design. Of course, you would probably want to follow a significant $F$ with pairwise tests on the difference scores to determine which groups differ significantly. Note that when the RM factor has only two levels, the sphericity assumption does not apply, and homogeneity of covariance across groups becomes the ordinary homogeneity of variance for the difference scores.

Although testing the interaction of the mixed design, or testing the difference scores, as described in the preceding, seem to be the most common ways of evaluating group differences in a before-after design, they are not the most powerful. A more sensitive test of group differences is based on a procedure known as the analysis of covariance (ANCOVA). Instead of simply subtracting the before score from the after score, you can use linear regression to predict the after score from the before score. The analysis proceeds...
in terms of the residual scores (after score minus predicted after score) rather than on the difference scores. The residual scores always have less variance (unless the regression slope happens to be 0 or 1.0) and therefore tend to yield a higher $t$ or $F$. The logic and mechanics of ANCOVA will be explained further in Chapter 18, Section B.

**Post Hoc Comparisons**

**When the Two Factors Do Not Interact**

If the interaction of the two factors is not statistically significant (nor large and disordinal), whichever main effects are significant can be explored in a straightforward manner. If the RM factor is significant and has more than two levels, a post hoc comparison method (e.g., Tukey’s HSD) can be used to test each pair of RM levels for significance (see Figure 16.6). If you feel strongly that all of the assumptions of the mixed design ANOVA have been met, $M_{S_{RM}}$ can be used as the error term for testing each pair of levels. However, if there is any doubt about the sphericity assumption, the safe thing to do is to base your error term on the variability of the difference scores for only the two levels being compared, for reasons discussed in the previous chapter. (The subject by treatment interaction for the two levels would be calculated separately for each group and then pooled, assuming that these interactions were similar from group to group.) Of course, complex comparisons can also be conducted among the RM levels using Scheffé’s test (for the preceding example, “before” vs. the average of “after” and “follow-up” would be a likely contrast followed by a test between “after” and “follow-up”). Again, it is safer to base your error term on only the RM levels involved in the comparison.

Similarly, a significant between-groups factor with more than two levels would be followed with pairwise or complex comparisons among the group means (averaging across the repeated measures), using the between-groups error term, $M_{S_w}$. With the same number of subjects in each group, homogeneity of variance across the groups is not a serious concern. However, if the groups have different sizes and considerably different variances, the equivalent of separate-variance $t$ tests should be used.

**Figure 16.6**

Pairwise Comparisons in a Mixed Design (When the Interaction Is Not Significant)

Pairwise comparisons between groups (collapsing across levels of the RM factor)

Pairwise comparisons between levels of the RM factor (collapsing across groups)

Pairwise comparisons following a mixed design with significant main effects, but no significant interaction between the factors.
predicted after score) cores always have less 1.0) and therefore as of ANCOVA will be

deficient (or large it can be explored in a short and has more than t's HSD) can be used gure 16.6). If you feel
ignoring ANOVA have been ng each pair of levels. assumption, the safeility of the differences ans discussed in the ion for the two levels hen pooled, assuming oup.) Of course, com-
t the RM levels using the average of “after” e by a test between ur error term on only 

An Excerpt from the Psychological Literature

The following example of a mixed design ANOVA in the psychological literature comes from an article entitled “Affective Valence and Memory in Depression: Dissociation of Recall and Fragment Completion” (Denny & Hunt, 1992). This study contains several ANOVAs, but the one I have chosen resembles the example in Section A, except that it does not include neutral words. Thus word valence refers to whether a word tended to evoke positive or negative affect. The results were reported as follows.

Recall data were subjected to an analysis of variance (ANOVA) with group as a between-subjects variable and word valence as a within-subjects variable. The results revealed a significant main effect for group: Recall level was higher for the nondepressed group than for the depressed group, $F(1, 30) = 30.21, MS_e = 2.05, p < .0001$. This effect was qualified, however, by a highly significant Group × Word Valence interaction, $F(1, 30) = 30.29, MS_e = 1.45, p < .0001$. The results of $t$-tests indicated that, as predicted, the depressed group recalled more negative than positive words, $t(15) = 4.45, SE_{difference} = .393, p < .001$. Within the nondepressed group, the opposite pattern was observed. Recall of positive words was significantly higher than that of negative words, $t(15) = 3.42, SE_{difference} = .456, p < .01$. Finally, comparisons revealed a significant between-groups difference in recall of positive words, $t(30) = 6.99, SE_m = .517, p < .001$, but not negative words, $t(30) = .75, SE_m = .411, p < .20$.

Note that the article gives the error term (MS_e) for each $F$ ratio after stating that ratio; this is the recommended practice so that a reader is equipped to perform his or her own follow-up analyses. Similarly, the denominator is given for the two-sample independent $t$ test (SE_m) and for the matched $t$ test (SE_{difference}). The authors refer to the interaction as being “highly significant.” Be aware that statistical purists abhor this expression; they argue that a result is either significant (i.e., $p$ is less than the predetermined alpha) or it is not. What the authors meant in this case is that the
Chapter 16 • Two-Way Mixed Design ANOVA

The order for calculating the SS components for the mixed design, as given in step 5 of this section, is meant to be instructive and reveal the concepts behind the analysis. A simpler, but equivalent, description follows. All SS components not found by subtraction are found by multiplying the biased variance of the appropriate means by the total number of observations (i.e., the number of different subjects or blocks times the number of levels of the repeated measures factor).

1. First, calculate the basic SS components of an independent two-way ANOVA (as in Chapter 14): $SS_{\text{total}}$, as usual, is based on all of the individual scores, and $SS_{\text{between-cell}}$ is based on the cell means (i.e., the mean for each group at each level of the RM factor); $SS_{\text{column}}$ is called $SS_{\text{RM}}$, and $SS_{\text{row}}$ is called $SS_{\text{between}}$ (assuming the row and column arrangement of Chapter 15). An additional component, $SS_{\text{group}}$, is found from the means of the groups (averaging across all subjects in the group and all levels of the RM factor). The remaining SS components can be found by subtraction.

2. Divide $SS_{\text{RM}}$, $SS_{\text{group}}$, $SS_{\text{RM} \times \text{factor}}$, and $SS_{\text{cell}}$ by their respective dfs to create the needed $MS$s.

3. If the different independent groups have different numbers of subjects, or if for every level of the between-subjects factor. This assumption can be tested with Box’s $M$ criterion. If $M$ is significant, the mixed design ANOVA, as presented in this chapter, is questionable. If $M$ is not significant, a sphericity test should be performed after pooling the pairwise interactions across groups.

4. If the sphericity test is significant, the df should be adjusted for both the RM factor and the two-way interaction before finding the critical $F$ in each case. Even if the sphericity test is not significant, an adjustment should be considered when your sample sizes are small.

5. If two independent groups are measured before and after some treatment that differs for the two groups, it is usually only the interaction of the mixed design that is of interest, and this $F$ ratio can also be found by performing an independent $t$ test on the before-after difference scores of the two groups and squaring the result. Similarly, if there are
more than two independent groups, a one-way ANOVA can be performed on the difference scores. However, an analysis of covariance of the "after" scores, using the "before" scores as the covariate is almost always the more powerful way to analyze this design.

6. When the two-way interaction is not significant, either main effect can be followed with post hoc comparisons in a routine way if it is significant. However, although there is little concern about using $MS_w$ from the overall analysis to test pairs of groups, $MS_{S \times RM}$ should only be used for pairs of RM levels if there is a strong reason to believe that sphericity exists in the population. When in doubt, your error term should be based on only the two RM levels being compared.

---

1. Imagine a study conducted to compare the effects of three types of training on the acquisition of a motor skill. Thirty-six subjects are divided equally into three groups, each receiving a different type of training. The performance of each subject is measured at five points during the training process.

   a. Construct a df tree (like the one in Figure 16.4) for the mixed design ANOVA that would be used to analyze this experiment.

   b. Find the critical values of $F$ for each of the $F$ ratios that would be calculated.

2. In Exercise 14B2, a group of visualizers and a group of nonvisualizers were each divided in half, and all subjects were presented with either concrete or abstract words to recall. A more powerful way to conduct that experiment would be to give each subject the same list containing a mixture of concrete and abstract words. The data from Exercise 14B2 follow, rearranged to make it easier to see the number of concrete and abstract words each subject recalled.

<table>
<thead>
<tr>
<th>Concrete</th>
<th>Abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visualizers</td>
<td>17</td>
</tr>
<tr>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>21</td>
<td>17</td>
</tr>
<tr>
<td>20</td>
<td>16</td>
</tr>
<tr>
<td>Nonvisualizers</td>
<td>18</td>
</tr>
<tr>
<td>19</td>
<td>18</td>
</tr>
<tr>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>20</td>
<td>19</td>
</tr>
</tbody>
</table>

a. Perform a two-way mixed design ANOVA on the data above (Note: You can save time by using some of the $SS_C$ that you calculated for Exercise 14B2.)

b. Present your results in a summary table and compare it to the summary table you created for Exercise 14B2.

3. A psychologist is studying the relationship between emotion and eating. All of his subjects view the same two film segments. One segment evokes happiness and one segment evokes fear; the order in which subjects view the film segments is counterbalanced. Half of the subjects are randomly assigned to a condition that requires them to eat a full meal just before viewing the film segments (preload condition); the remaining half are not permitted to eat during the 4 hours preceding the experiment (no load condition). The subjects are offered an unlimited amount of popcorn while viewing the film segments. The amount of popcorn (in ounces) consumed by each subject in each condition appears in the following table.

<table>
<thead>
<tr>
<th>Happiness</th>
<th>Fear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preload</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>No Load</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>27</td>
<td>20</td>
</tr>
<tr>
<td>20</td>
<td>15</td>
</tr>
</tbody>
</table>

a. Perform a mixed design ANOVA.

b. Draw a graph of the cell means. (Note that you already graphed the data for the individual subjects in Exercise 16A4.) Describe the nature of each significant effect.
c. Calculate the happiness-fear difference score for each subject, and then perform a two-group independent t test on these difference scores. Which of the F ratios calculated in part a is related to the t value you just found? What is the relationship?

4. In Exercise 15B1, subjects performed a clerical task under three noise conditions. Now suppose a new group of subjects is added to study the effects of the same three conditions on the performance of a simpler, more mechanical task. The data from Exercise 15B1 follow, along with the data for the mechanical task. Perform a mixed design ANOVA, and display the results in a summary table.

<table>
<thead>
<tr>
<th>Background</th>
<th>Popular</th>
<th>Heavy Metal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clerical Task</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Mechanical Task</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>19</td>
<td>22</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>16</td>
<td>19</td>
<td>19</td>
</tr>
</tbody>
</table>

5. Dr. Jones is investigating various conditions that affect mental effort—which, in this experiment, involves solving anagrams. Subjects were randomly assigned to one of three experimental conditions. Subjects in the first group were told that they would not be getting feedback on their performance. Subjects in the second and third groups were told they would get feedback, but only subjects in the third group were told (erroneously) that anagram solving was highly correlated with intelligence and creativity (Dr. Jones hoped this information would produce ego involvement). The list of anagrams given to each subject contained a random mix of problems at four levels of difficulty determined by the number of letters presented (five, six, seven, or eight). The number of anagrams correctly solved by each subject in each condition and at each level of difficulty is given in the following table:

<table>
<thead>
<tr>
<th>Blocks</th>
<th>Control</th>
<th>Biofeedback</th>
<th>Drug</th>
<th>Self-Hypnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man</td>
<td>14</td>
<td>5</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>9</td>
<td>7</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>12</td>
<td>3</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>15</td>
<td>7</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>8</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>14</td>
<td>4</td>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>

a. Before conducting the ANOVA, graph the cell means and guess which of the three F ratios is (are) likely to be significant and which is (are) not.

b. Test the significance of each of the three F ratios.

c. Describe the post hoc pairwise comparisons that would be appropriate for this experiment, given your answer to part b.

7. A market researcher is comparing three types of commercials to determine which will have the largest positive effect on the typical consumer. One type of commercial is purely informative, one features a celebrity
endorsement, and the third emphasizes the glamour and style of the product. Six different subjects are randomly assigned to watch each of the three types of commercials. Subjects rate their likelihood of buying the product on a scale from 0 (very unlikely to buy) to 10 (very likely to buy) both before and after watching the assigned commercials. The data appear in the following table:

<table>
<thead>
<tr>
<th>INFORMATIVE</th>
<th>CELEBRITY</th>
<th>GLAMOUR/STYLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>After</td>
<td>Before</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

a. Perform a mixed design ANOVA and test the three F ratios at the .01 level.
b. Calculate the before-after difference score for each subject, and then perform a one-way independent groups ANOVA on these difference scores. Which of the F ratios calculated in part a is the same as the F ratio you just found? Explain the connection.

8. Exercise 15B6 described a neuropsychologist studying subjects with brain damage to the left cerebral hemisphere. Such a study would probably include a group of subjects with damage to the right hemisphere and a group of control subjects without brain damage. The data from Exercise 15B6 (the number of digit or letter strings each subject recalled) follow, along with data for the two comparison groups just mentioned.

<table>
<thead>
<tr>
<th>Digit</th>
<th>Letter</th>
<th>Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left Brain Damage</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Right Brain Damage</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Control</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>9</td>
</tr>
</tbody>
</table>

9. Perform all of the possible pairwise comparisons for the between-groups factor in Exercise 5, averaging across the RM levels. Do you need to adjust alpha for each comparison to protect the experimentwise Type I error rate?

10. Perform all of the possible pairwise comparisons for the RM factor in Exercise 6, averaging across gender.

C

**Post Hoc Comparisons When the Two Factors Interact**

If $F_{G \times RM}$ is statistically significant, the main effects are usually not explored further, and post hoc comparisons are focused instead on the cell means, along the lines already described in Chapter 14, Section C. The most common approach is an analysis of simple main effects. You can test the effects of the RM factor separately for each group or test the effects of the between-groups factor separately at each level of the RM factor or conduct...
tests in both ways (see Figure 16.8). Unfortunately, the interaction of the two factors in a mixed design complicates the choice of an error term for each of the simple effects, as I will explain.

The simplest and safest way to test simple effects involving the RM factor is to conduct a separate one-way RM ANOVA for each group, as though the other groups do not exist. If there is homogeneity among the groups in terms of the subject by treatment interaction, more power can be gained by using $MS_{S \times RM}$ from the overall analysis rather than the $MS_{S \times RM}$ for just the group being tested. However, the more conservative approach, in light of a significant interaction between the factors, is to sacrifice the small amount of extra power from the pooled error term, in favor of not increasing the risk of a Type I error, in case there actually is no homogeneity among the groups in the population.

To follow up a significant simple effect, you would probably want to conduct pairwise comparisons among the different RM levels (e.g., before versus after for the control group; after versus follow-up for the experimental group). If you had a strong reason to assume sphericity within the significant simple effect, you could use the $MS_{S \times RM}$ for that group as the error term for each pairwise test. However, because the sphericity assumption is generally considered quite risky, especially for pairwise comparisons, it is strongly recommended that you base your error term only on the two levels being tested. (This is equivalent to performing a simple matched $t$ test between a pair of RM levels for one of the groups.) The Type I error rate can be controlled by using the Bonferroni test, setting alpha at .05 for the entire family of possible RM comparisons.

Simple effects can also be tested by comparing the different groups at each level of the RM factor. However, the proper error term for each of these effects is not $MS_w$, as calculated for the mixed design. You may have noticed that what I'm calling $MS_w$ in the mixed design is not the same as $MS_w$ in a two-way independent ANOVA. In the independent ANOVA, variances are calculated for each cell and then averaged (I'll call this $MS_{within-cell}$). In the mixed design, variances are not calculated separately for each cell; scores are averaged over the RM levels, and then variances

---

**Figure 16.8**

Pairwise Comparisons in a Mixed Design (When the Interaction Is Significant)

---

Pairwise comparisons for a significant simple effect of the between-groups factor

---

Pairwise comparisons for a significant simple effect of the RM factor

---

Pairwise comparisons following a mixed design with a significant interaction
are calculated within each group (I'll call MS_{within-group}. In fact, if you calculated SS_{within-cell} for a mixed design, it would equal the sum of SS_{within-group} and SS_{x \times RM}.

The relation just described provides another way to understand the error terms in the mixed design. You can begin by calculating SS_{within-cell} as for an independent ANOVA. Then this SS is divided into two pieces: one ignores subjects' differing reactions to different RM levels by averaging across those levels (SS_{within-group}), whereas the other piece is insensitive to overall differences between subjects and measures only the extent to which subjects react similarly to the differing RM levels within each group (SS_{x \times RM}). If we are performing post hoc comparisons in which we are looking at differences among the groups at one level of the RM factor at a time, it is not appropriate to use an error term that averages across RM levels (i.e., MS_{within-group}), especially given that the grouping variable interacts significantly with the RM factor. The appropriate error term is MS_{within-cell}.

Unfortunately, the homogeneity assumption required to justify the use of MS_{within-cell} as the error term for post hoc comparisons is likely to be violated, leading to a biased F ratio. For large sample sizes this bias will be negligible, but for samples smaller than 30 an adjustment of the degrees of freedom (similar to the adjustment in the separate-variances t test) is recommended before you determine the critical F. The problem can be avoided, with some loss of power, by pooling error terms for only the cells involved in the analysis. For instance, if you are comparing only “after” measurements for two experimental groups and a control group, the error term can be based on pooling the MS_w just those cells, rather than being based on MS_{within-cell} from the entire mixed design. This localized error term can then be used to test pairs of groups within a particular level of the RM factor, whenever the simple effect of groups is significant at that level of the RM factor. (Of course, if there are only two groups in the design, there are no follow-up tests to be done on that factor.)

Planned and Complex Comparisons

As an alternative to simple effects analysis, interaction contrasts can be performed on a mixed design, either as planned tests (using a set of orthogonal contrasts or a Bonferroni correction) or as post hoc tests (using Scheffe’s test). For the example in Section B, one reasonable pair of orthogonal contrasts that might be planned is (1) “before” and the average of “after” and “follow-up” crossed with the two groups and (2) “after” and “follow-up” crossed with the two groups. The error term for (1) would be the same as the omnibus ANOVA, but it can be argued that the safe error term for (2) is one that does not involve the “before” scores.

For larger mixed designs (at least 3 \times 3), the partial interactions described in Chapter 14, Section C can be considered. However, although the same error term can almost always be used for all possible contrasts in a two-way independent groups design, concerns about the sphericity assumption require caution about the choice of error term in the mixed design, along the lines previously discussed for simple effects. One solution that is particularly simple when using statistical software is to calculate each partial interaction or interaction contrast as a separate mixed design ANOVA without attempting to use the error term from the larger analysis. Some power is sacrificed in favor of increased caution regarding the experimentwise alpha rate.
Removing Error Variance from Counterbalanced Designs

In a one-way repeated-measures experiment, counterbalancing prevents the presence of simple order effects from systematically affecting the numerator of the $F$ ratio, but order effects will inflate the denominator, which in turn reduces the power of the test. The extra variance produced by order effects can and, in most cases, should be removed from your error term before you obtain your $F$ ratio and test it for significance. The easiest way to do this is to convert your one-way RM ANOVA into a two-way mixed ANOVA in which each group consists of subjects who were given the RM levels in the same order. For instance, if your IV had four levels and you used a Latin square design, your mixed design would have four groups, each representing a different order (your total number of subjects would have to be divisible by 4). (I am describing the single Latin square design with replications. Instead of having just four orders, you could have a different Latin square for each set of four subjects. Such a design is considerably more tedious to run and to analyze; therefore, it is rarely used, even though it provides a more thorough balancing of possible complex order effects.) I will refer to the added between-groups factor as the "order" factor and to the RM factor as the "treatment" factor.

In the mixed design described in the preceding, both simple order effects and asymmetric carryover effects contribute to the order x treatment interaction. If it is not obvious that simple order effects would contribute to that interaction, it will help if you imagine a memory study in which each subject's recall is tested in two rooms, one painted red and one painted green. Imagine also that room color has no effect on memory, so each subject's line on a graph—with red and green on the horizontal axis—should be flat. However, if the red and green conditions are counterbalanced and there is a practice effect that gives a boost to whichever condition is second, the lines of the "red-first" subjects will slant one way, and the "green-first" subjects will slant the opposite way, creating an order x treatment interaction. Without adding order as a factor to the ANOVA, the order effect contributes to the subject by treatment interaction—(i.e., the error term for the one-way analysis). Adding the order factor separates the order x treatment interaction from the subject x treatment interaction, so the latter is no longer influenced by order effects (the SS for the subject x treatment interaction is calculated separately for each order and then summed).

Performing the mixed design described in the preceding tends to increase the $F$ ratio for your treatment effect when there are order effects embedded in your ANOVA. However, you can lose a considerable number of degrees of freedom in your error term by adding order as a factor, and this makes the error term larger ($MS_{error} = SS_{error}/df_{error}$), so unless you actually have considerable order effects, adding order can hurt more than help. If you perform the mixed design ANOVA and find that both the order main effect and its interaction with the treatment are very small, it is recommended that you then drop order as a factor and proceed with your analyses in the usual manner. If your original study is already a mixed design (e.g., three different cognitive tasks, with half the subjects told that the tasks measure intelligence and half that they do not), adding task order as a second between-subjects factor creates a three-way ANOVA with one RM factor.

If the order x treatment interaction is large in your mixed ANOVA, you know that you have large order effects that have been removed from your
error term. That sounds like a problem solved, but, unfortunately, both differential carryover effects and simple order effects can contribute to the order \times treatment interaction. You may recall from the previous chapter that differential carryover effects are a type of order effect that can represent a confounding variable in your study, spuriously increasing or decreasing the separation of your treatment means. Before you accept the conclusions of your mixed design, you would want to rule out the possibility of significant carryover effects. You can check for carryover effects by drawing a graph of your treatment means versus serial position as in Figure 15.8. The further the lines on your graph deviate from being parallel, the larger the carryover effects. To quantify the magnitude of your carryover effects, you can use the following analysis.

The analysis I'm about to describe is usually referred to as the analysis of a Latin square design, although it would work as well for a design that is completely counterbalanced (the latter design is rarely used when there are more than three treatment levels). The Latin square analysis begins with the mixed design I described in the preceding but further refines it by dividing the order \times treatment effect into two pieces. I mentioned that one problem with the mixed design approach is that the order \times treatment interaction contains both simple order and carryover effects. To separate these two components, it is necessary to quantify the simple order effect by calculating the SS for the main effect of serial position (SS_p). This is found by calculating the mean of your dependent variable at each serial position (e.g., to find the mean for position 1, you have to average every subject's first condition together even though you will be averaging together different experimental conditions). For example, using the orders in Table 15.4, I have rearranged the data in Table 15.5 in terms of serial position instead of the actual experimental conditions and created Table 16.7. You can see by the column means in Table 16.7 that there is a practice effect such that tasks performed later have an advantage.

<table>
<thead>
<tr>
<th>Subject No.</th>
<th>Position 1</th>
<th>Position 2</th>
<th>Position 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>20</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>7</td>
<td>18</td>
</tr>
<tr>
<td><strong>Column Means</strong></td>
<td><strong>10.833</strong></td>
<td><strong>13.167</strong></td>
<td><strong>14.167</strong></td>
</tr>
</tbody>
</table>

SS_p is calculated as for any main effect: N_r \cdot \sigma^2(position means). Then SS_p is subtracted from the order by treatment interaction; the leftover portion is the position by treatment interaction (SS_p \times \tau), which is a measure of the variability due to differential carryover effects. SS_p \times \tau can be tested for significance by first dividing it by its df, which equals (c - 1)(c - 2), to create MS_p \times \tau and then dividing by the MS_p \times RM error term from the mixed design ANOVA. This is one F ratio that you do not want to be significant or even close to significance. A significant position by treatment interaction, together with a graph showing noticeable carryover effects, suggests that you should reanalyze your data using only the first treatment condition for each subject. This converts your RM design to a between-subjects design, which means a large loss of data and power, but any significant results you do get cannot be attributed to carryover effects. Finding that the results of
the between-group analysis are similar to the RM analysis is often used as justification for retaining the RM analysis, but any serious discrepancy argues for using the between-group results.

You can also test $MS_{p}$ (divide $SS_{p}$ by $c - 1$) for significance, using the same error term as for $MS_{px\gamma}$. The significance of this effect won’t affect your interpretation of your main treatment effect, but it can be interesting in its own right. A significant main effect of position may demonstrate the importance of a practice effect, a fatigue effect, or both (the means may be highest for the middle positions). If you are wondering how to do the Latin square analysis with one of the major statistical packages, there is no easy way at present. Probably the simplest choice is to use the computer to do the mixed design analysis and then calculate $SS_{p}$ with a handheld calculator and complete the analysis yourself. You can get the computer to calculate $SS_{p}$, but, unless your data set is huge, it hardly seems worth the trouble because you would have to either reenter the data by position instead of condition or create new position variables with program statements or a series of menu operations.

The Latin square analysis just described assumes that you have more than one subject for each set of orders; otherwise you could not perform a mixed design. But even if each subject in your experiment receives a different order of experimental conditions (e.g., there are eight conditions and eight subjects run in a Latin square design), it is still possible to remove order effects from your error term and test the main effect of position (but no longer possible to test the position by treatment interaction).

The example in Section B of the previous chapter is not a Latin square (it is completely counterbalanced), but it involves only one subject per order. To extract order effects from that design, begin with the usual RM analysis, as performed in the previous chapter. To calculate $SS_{p}$, it is helpful to rearrange the scores according to their serial positions, as I did in Table 16.7. Then, $SS_{p} = 18 \cdot \sigma^{2}(10.83, 13.167, 14.167) = 18 \cdot 1.951 = 35.12$. When there is only one subject per order, the order \times treatment interaction is identical with the subject \times treatment interaction, so $SS_{p}$ is subtracted from $SS_{x \times RM}$ to create an adjusted error term: $SS_{adj\_error} = 69.2 - 35.12 = 34.08$. $SS_{adj\_error}$ must be divided by the proper df to create $MS_{adj\_error}$. The original $df_{error}$ was $2 \cdot 5 = 10$. Extracting $SS_{p}$, however, results in the loss of two df (the number of df lost is always one less than the number of levels or serial positions), so $df_{adj\_error} = 8$. $MS_{adj\_error} = SS_{adj\_error}/8 = 34.08/8 = 4.26$. $MS_{adj}$ is still 55.4, so the new $F$ ratio is $55.4/4.26 = 10.53$, which is considerably larger than the original $F$ ratio of 8.0. The reduction in the $SS$ for the error term outweighed the reduction in $df_{error}$, which is generally the case when the order effects are fairly large.

You can also test the main effect of serial position in the design just described. If you divide $SS_{p}$ by $df_{p}$, you get $MS_{p}$, which equals $35.12/2 = 17.56$. Then divide $MS_{p}$ by the same adjusted error term I just calculated: $F_{p} = 17.56/4.26 = 4.12$. The critical $F_{crit}(2, 8) = 4.46$, so the serial position effect falls short of significance. This is no surprise with such a small sample size, and it seems reasonable to make the error adjustment even though the order effect fails to reach significance. The adjusted error term is a measure of the position by treatment interaction, but it cannot be tested for significance (there is no additional error term to test it against). Therefore, with one subject per order, one must look carefully for carryover effects before making conclusions. Also, if your design is a Latin square with one subject per order, the loss in degrees of freedom due to extracting $SS_{p}$ can reduce your power to a ridiculously low level. If you are going to run such a design, the just described analysis is not recommended unless you have at least five conditions/subjects.
Relative Efficiency

The advantage of extracting $SS_p$ can be expressed in terms of the relative efficiency (RE) of that analysis as compared to the ordinary one-way RM ANOVA. The RE for two corresponding analyses can be expressed as a ratio of the error terms of the two analyses, as shown in the following formula:

$$RE = \frac{MS_{error1}}{MS_{error2}} \left( \frac{df_{error1} + 1}{df_{error1} + 3} \right) \left( \frac{df_{error2} + 3}{df_{error2} + 1} \right)$$  \hspace{1cm} \text{Formula 16.8}

If the analysis subtracting $SS_p$ is considered analysis 1, and the original analysis is considered analysis 2, the relative efficiency of the two analyses can be found by inserting the relevant values into the preceding formula:

$$RE = \frac{6.92}{4.26} \left( \frac{8 + 1}{8 + 3} \right) \left( \frac{10 + 3}{10 + 1} \right) = (1.624)(.818)(1.18) = 1.57$$

Expressed as a percentage, the relative efficiency of the $SS_p$ analysis is 157% as compared to the ordinary RM ANOVA, which makes it well worth the trouble. Another interesting use of the RE formula is to quantify the advantage of an RM ANOVA relative to its between-groups counterpart (an interesting exercise would be to do this for the preceding analysis). And if you compare the $SS_p$ analysis directly to the analysis that doesn’t use repeated measures at all, the RE should be quite impressive (it is more than 270% for the preceding example).

1. A significant interaction is often followed by an analysis of simple (main) effects. Simple effects of the RM factor are most safely explored by conducting separate one-way RM ANOVAs for each group; $MS_{x \times RM}$ from the overall analysis should be used as the error term for simple effects only when there is a strong basis to assume that all of the homogeneity assumptions of the mixed design have been met. Similarly, significant simple RM effects can be followed by pairwise tests for pairs of levels, but the error term for the pairwise tests should be based on the two levels involved, rather than the pooled error term from the simple effect, unless sphericity can be assumed for that simple effect.

2. Simple effects can also be tested among the groups at each level of the RM factor. For these tests, the appropriate error term is $MS_{within-cell}$, the average variance of the cells, as in a two-way independent ANOVA (in a mixed design, $SS_{within-cell}$ is subdivided into $SS_{x \times RM}$ and a portion I’ve been calling $SS_p$ to form the basis of the two error terms). For small samples, and when the homogeneity of cell variances is in doubt, the safer approach is to pool the variances for the cells involved in the simple effect tested. This more limited pooled error term can then be used to test pairs of groups for a simple effect that reached significance.

3. As with any two-way ANOVA, a significant interaction can be followed by a test of partial interactions or interaction contrasts to further localize the effect (such tests can also be planned). However, unlike the independent-groups ANOVA, in which $MS_p$ is almost always used as the error term for all comparisons, sphericity concerns argue against using $MS_{x \times RM}$ from the overall analysis in the same way. The simple solution is to perform a separate mixed design ANOVA on some subset of the full design (e.g., interaction contrast), using the error term from that subset.

4. In a counterbalanced design, order effects contribute to the subject by treatment interaction and therefore make the error term for the RM
factor larger. Separating subjects according to the order in which they received the treatments and adding “order” as a between-groups factor removes order effects from the subject by treatment interaction and places them in a group by treatment interaction, where they can be tested for significance. Removing order effects often increases the F ratio of the RM factor, but adding order as a factor always reduces the df of the error term, which tends to reduce the F ratio. If order effects are small, the increase in F due to removing order effects won’t offset the decrease in F due to the decrease in df, so the order factor should not be added in such cases.

5. If your order by treatment interaction is significant and/or very large and disordinal, you may want to separate simple order effects from complex effects because the latter may threaten the validity of your results. The mean can be calculated for each ordinal position (across treatments). Then an SS can be calculated from these means and subtracted from the SS for the order x treatment interaction. If the order x treatment interaction is still significant, you should inspect a graph of treatment by position to look for differential carryover effects that may be confounded with your main RM effect. If strong complex order effects are evident, you may have to convert your RM design to a between-groups design by including only the first treatment given to each subject.

6. The preceding analysis is usually performed on the results of a single Latin square design, in which the number of orders equals the number of treatments, and the number of subjects is some multiple of the number of treatments (it can also be applied to a case of complete counterbalancing). If your design is a Latin square (or completely counterbalanced design) with only one subject per order, you can still calculate the SS for the main effect of position and subtract order effects from the error term, but you cannot separate simple from complex order effects or test the latter separately. You must rely instead on a position by treatment graph to determine if your order effects are simple or potentially confounding.

7. Adding order as a factor to an RM design creates two changes that have opposite effects on the F ratio and hence on power. The error term tends to go down, but so do the df for the error term. These two opposing effects are incorporated in a measure called the relative efficiency (RE) of two analyses. If the two analyses differ only in their error terms and corresponding df’s, RE gives us a measure of the advantage one design or analysis has over the other. The RE can give you some idea of whether matching subjects was worth the trouble for a particular study or whether adding order as a factor makes a large difference.

---

1. Given that the interaction was significant for Exercise 3 in Section B, describe all of the post hoc comparisons that would be justified. Perform those tests, using the conservative error term in each case.

2. For Exercise 4 in Section B, perform a complete analysis of the simple effects, including pairwise comparisons for significant simple effects. Use the appropriate omnibus error term in each case, given that the interaction for the full design is significant.

3. For Exercise 8 in Section B
   a. Describe a meaningful partial interaction as defined in Chapter 14, Section C. What
follow-up tests would be appropriate if your partial interaction turned out to be significant?

b. Describe a complete set of orthogonal contrasts that answer meaningful questions about the data. To be conservative, how would you find the error term for each of your contrasts?

*4. Suppose that the experiment described in Exercise 15B3 was not completely counterbalanced and that the first four subjects followed the order "One," "Twenty," "Large,"; the second four followed "Twenty," "Large," "One,"; and the last four, "Large," "One," "Twenty."

a. Add order as a between-groups factor, and complete the mixed design ANOVA. Compare the F ratio for the RM factor in the mixed design with the F ratio from the original one-way RM ANOVA.

b. Draw a graph of the treatments as a function of serial position. Can you see evidence of complex order effects?

c. Complete the Latin square analysis. Test the main effect of position and the position by treatment interaction for significance.

5. In Exercise 15B6, the orders in which the subjects received the Digit (D), Letter (L), and Mixed (M) tasks were as follows:
   1. DLM; 2. DML; 3. LMD; 4. LDM; 5. MDL; 6. MLD.

a. Complete the Latin square analysis, and draw a graph of the treatments as a function of serial position. Is the main effect of position significant?

b. Compare the F ratio for the RM factor in the Latin square design, with the F ratio from the original one-way RM ANOVA.

*6. a. Calculate the relative efficiency (RE) of the mixed design in Exercise 4 in comparison with the original one-way RM ANOVA.

b. Calculate the (RE) of the Latin square design in Exercise 5 in comparison with the original one-way RM ANOVA.

The df components for the mixed design.

\[ \text{df}_{\text{between-S}} = nk - 1 \]

\[ \text{df}_{\text{within-S}} = nk(c - 1) \quad \text{or} \quad nkc - nk \]

\[ \text{df}_{\text{groups}} = k - 1 \]

\[ \text{df}_{\text{w}} = k(n - 1) \quad \text{or} \quad nk - k \]

\[ \text{df}_{\text{RM}} = c - 1 \]

\[ \text{df}_{G \times \text{RM}} = (k - 1)(c - 1) \]

\[ \text{df}_{S \times \text{RM}} = k(c - 1)(n - 1) \]

The sum of squares due to subject-to-subject variability within groups, found by subtraction:

\[ SS_w = SS_{\text{between-S}} - SS_{\text{groups}} \]

The sum of squares due to variation among the several measurements within each subject (this component is further divided into subcomponents corresponding to mean differences between levels of the RM factor, the interaction of the two factors, and the interaction of the subjects with the RM factor within each group):

\[ SS_{\text{within-S}} = SS_{\text{total}} - SS_{\text{between-S}} \]

The sum of squares due to the interaction of the two factors, found by subtraction:

\[ SS_{G \times \text{RM}} = SS_{\text{between-cells}} - SS_{\text{groups}} - SS_{\text{RM}} \]