CHAPTER FOURTEEN

REPEATED-MEASURES DESIGNS

OBJECTIVES

To discuss the analysis of variance by considering experimental designs in which the same subject is measured under all levels of one or more independent variables.

CONTENTS

14.11. OTHER CONSIDERATIONS

14.12. MIXED MODELS FOR REPEATED MEASURES DESIGNS

In our discussion of the analysis of variance, we have concerned ourselves with experimental designs that have different subjects in the different cells. More precisely, we have been concerned with designs in which the cells are independent, or uncorrelated. (Under the assumptions of the analysis of variance, *independent* and *uncorrelated* are synonymous in this context.) In this chapter we are going to be concerned with the problem of analyzing data where some or all of the cells are not independent. Such designs are somewhat more complicated to analyze, and the formulae because more complex. Most, or perhaps even all, readers will approach the problem using computer software such as SPSS or SAS. However, to understand what you are seeing, you need to know something about how you would approach the problem by hand; and that leads to lots and lots of formulae. I urge you to treat the formulae lightly, and not feel that you have to memorize any of them. This chapter needs to be complete, and that means we have to go into the analysis at some depth, but remember that you can always come back to the formulae when you need them, and don't worry about the calculations too much until you do need them.

If you think of a typical one-way analysis of variance with different subjects serving under the different treatments, you would probably be willing to concede that the correlations between treatments 1 and 2, 1 and 3, and 2 and 3 have an expectation of zero.

Treatment 1	Treatment 2	Treatment 3
X_{11}	X_{21}	X_{31}
X_{12}	X_{22}	X_{32}
...	...	...
X_{1n}	X_{2n}	X_{3n}

However, suppose that in the design diagrammed here the same subjects were used in all three treatments. Thus, instead of 3*n* subjects measured once, we have *n* subjects measured three times. In this case, we would be hard put to believe that the intercorrelations of the three treatments would have expectancies of zero. On the contrary, the better subjects under treatment 1 would probably also perform well under treatments 2 and 3, and the poorer subjects under treatment 1 would probably perform poorly under the other conditions, leading to significant correlations among treatments.

This lack of independence among the treatments would cause a serious problem if it were not for the fact that we can separate out, or **partition**, and remove the dependence imposed by repeated measurements on the same subjects. (To use a term that will become much more familiar in Chapter 15, we can say that we are **partialling out** effects that cause the dependence.) In fact, one of the main advantages of **repeated-measures designs** is that they allow us to reduce overall variability by using a common subject pool for all treatments, and at the same time allow us to remove subject differences from our error term, leaving the error components independent from treatment to treatment or cell to cell.

As an illustration, consider the highly exaggerated set of data on four subjects over three treatments presented in Table 14.1. Here the dependent variable is the number of trials to criterion on some task. If you look first at the treatment means, you will see some slight

differences, but nothing to get too excited about. There is so much variability within each treatment that it would at first appear that the means differ only by chance. But look at the subject means. It is apparent that subject 1 learns quickly under all conditions, and that subjects 3 and 4 learn remarkably slowly. These differences among the subjects are producing most of the differences *within* treatments, and yet they have nothing to do with the treatment effect. If we could remove these subject differences we would have a better (and smaller) estimate of error. At the same time, it is the subject differences that are creating the high positive intercorrelations among the treatments, and these too we will partial out by forming a separate term for subjects.

Table 14.1 Hypothetical data for simple repeated-measures designs

	Treatment			
Subject	**1**	**2**	**3**	**Mean**
1	2	4	7	4.33
2	10	12	13	11.67
3	22	29	30	27.00
4	30	31	34	31.67
Mean	16	19	21	18.67

One laborious way to do this would be to put all the subjects' contributions on a common footing by equating subject means without altering the relationships among the scores obtained by that particular subject. Thus, we could set $X'_{ij} = X_{ij} - \bar{X}_i$, where $\bar{X}_i$ is the mean of the *i*th subject. Now subjects would all have the same means ($\bar{X}'_{i.} = 0$), and any remaining differences among the scores could be attributable only to error or to treatments. Although this approach would work, it is not practical. An alternative, and easier, approach is to calculate a sum of squares between subjects (denoted as either $SS_{\text{between subj}}$ or SS_s) and remove this from SS_{total} before we

begin. This can be shown to be algebraically equivalent to the first procedure and is essentially the approach we will adopt.

The solution is represented diagrammatically in Figure 14.1. Here we partition the overall variation into variation between subjects and variation within subjects. We do the same with the degrees of freedom. Some of the variation within a subject is attributable to the fact that his scores come from different treatments, and some is attributable to error; this further partitioning of variation is shown in the third line of the figure. We will always think of a repeated-measures analysis as *first* partitioning the SS_{total} into $SS_{between subj}$ and $SS_{within subj}$. Depending on the complexity of the design, one or both of these partitions may then be further partitioned.

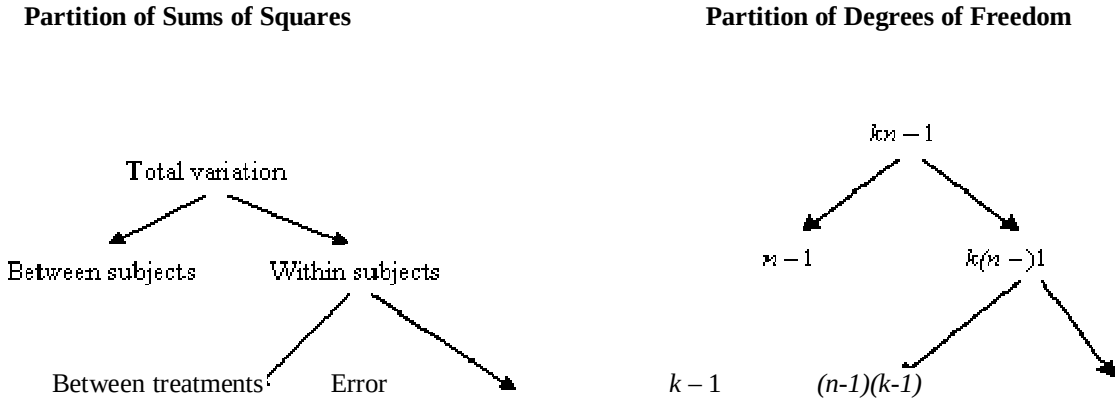


Figure 14.1 Partition of sums of squares and degrees of freedom

The following discussion of repeated-measures designs can only begin to explore the area. For historical reasons, the statistical literature has underemphasized the importance of these designs. As a result, they have been developed mostly by social scientists, particularly psychologists. By

far the most complete coverage of these designs is found in Winer, Brown, and Michels (1991). Their treatment of repeated-measures designs is excellent and extensive, and much of this chapter reflects the influence of Winer's work.

14.1. THE STRUCTURAL MODEL

First, some theory to keep me happy. Two structural models could underlie the analysis of data like those shown in Table 14.1. The simplest model is

$$X_{ij} = \mu + \pi_i + \tau_j + e_{ij}$$

where

μ = the grand mean
π_i = a constant associated with the ith person or subject, representing how much that person differs from the average person
τ_j = a constant associated with the jth treatment, representing how much that treatment mean differs from the average treatment mean
e_{ij} = the experimental error associated with the ith subject under the jth treatment

The variables π_i and e_{ij} are assumed to be independently and normally distributed around zero within each treatment. Their variances, σ^2_π and σ^2_ε, are assumed to be homogeneous across treatments. (In presenting expected means square, I am using the notation developed in the preceding chapters. The error term and subject factor are considered to be random, so those variances are presented as σ^2_π and σ^2_ε. (Subjects are always treated as random.) However the treatment factor is generally a fixed factor, so its variation is denoted as θ^2_τ) With these assumptions it is possible to derive the expected mean squares shown in Model I of Table 14.2.

Table 14.2 Expected mean squares for simple repeated-measures designs

Model I

$X_{ij} = \mu + \pi_i + \tau_j + e_{ij}$

Source	*E*(MS)
Subjects	$\sigma_e^2 + k\sigma_\pi^2$
Treatments	$\sigma_e^2 + n\theta_\tau^2$
Error	σ_e^2

Model II

$X_{ij} = \mu + \pi_i + \tau_j + \pi\tau_{ij} + e_{ij}$

Source	*E*(MS)
Subjects	$\sigma_e^2 + k\sigma_\pi^2$
Treatments	$\sigma_e^2 + \sigma_{\pi\tau}^2 + n\theta_\tau^2$
Error	$\sigma_e^2 + \sigma_{\pi\tau}^2$

An alternative and probably more realistic model is given by

$$X_{ij} = \mu + \pi_i + \tau_j + \pi\tau_{ij} + e_{ij}$$

Here we have added a Subject × Treatment interaction term to the model, which allows different subjects to change differently over treatments. The assumptions of the first model will continue to hold, and we will also assume the $\pi\tau_{ij}$ to be distributed around zero independently of the other elements of the model. This second model gives rise to the expected mean squares shown in Model II of Table 14.2.

The discussion of these two models and their expected mean squares may look as if it is designed to bury the solution to a practical problem (comparing a set of means) under a mountain of statistical theory. However, it is important to an explanation of how we will run our analyses and where our tests come from. You'll need to bear with me only a little longer.

14.2. *F* RATIOS

The expected mean squares in Table 14.2 indicate that the model we adopt influences the *F* ratios we employ. If we are willing to assume that there is no Subject × Treatment interaction, we can form the following ratios:

$$\frac{E(\mathrm{MS}_{\text{between subj}})}{E(\mathrm{MS}_{\text{error}})} = \frac{\sigma_\varepsilon^2 + k\sigma_\pi^2}{\sigma_\varepsilon^2}$$

and

$$\frac{E(\mathrm{MS}_{\text{treat}})}{E(\mathrm{MS}_{\text{error}})} = \frac{\sigma_\varepsilon^2 + n\theta_\tau^2}{\sigma_\varepsilon^2}$$

Given an additional assumption about sphericity, which we will discuss in the next section, both of these lead to respectable *F* ratios that can be used to test the relevant null hypotheses.

Usually, however, we are cautious about assuming that there is no Subject × Treatment interaction. In much of our research it seems more reasonable to assume that different subjects will respond differently to different treatments, especially when those "treatments" correspond to phases of an ongoing experiment. As a result we usually prefer to work with the more complete model.

The full model (which includes the interaction term) leads to the following ratios:

$$\frac{E(\mathrm{MS}_{\text{between subj}})}{E(\mathrm{MS}_{\text{error}})} = \frac{\sigma_\varepsilon^2 + k\sigma_\pi^2}{\sigma_\varepsilon^2 + \sigma_{\pi\tau}^2}$$

and

$$\frac{E(\mathrm{MS}_{\text{treat}})}{E(\mathrm{MS}_{\text{error}})} = \frac{\sigma_\varepsilon^2 + \sigma_{\pi\tau}^2 + n\theta_\tau^2}{\sigma_\varepsilon^2 + \sigma_{\pi\tau}^2}$$

Although the resulting *F* for treatments is appropriate, the *F* for subjects is biased. If we did form this latter ratio and obtained a significant *F*, we would be fairly confident that subject differences really did exist. However, if the *F* were not significant, the interpretation would be ambiguous. A nonsignificant *F* could mean either that $k\sigma_\pi^2 = 0$ or that $k\sigma_\pi^2 > 0$ but $\leq \sigma_{\pi\tau}^2$. Because we usually prefer this second model, and hate ambiguity, we seldom test the effect due to Subjects. This represents no great loss, however, since we have little to gain by testing the Subject effect. The main reason for obtaining $\mathrm{SS}_{\text{betweensubj}}$ in the first place is to absorb the correlations between treatments and thereby remove subject differences from the error term. A test on the Subject effect, if it were significant, would merely indicate that people are different—hardly a momentous finding. The important thing is that both underlying models show that we can use $\mathrm{MS}_{\text{error}}$ as the denominator to test the effect of treatments.

14.3. THE COVARIANCE MATRIX

A very important assumption that is required for any *F* ratio in a repeated-measures design to be distributed as the central (tabled) *F* is that of compound symmetry of the covariance matrix.[1] To understand what this means, consider a matrix ($\hat{\Sigma}$) representing the covariances among the three treatments for the data given in Table 14.1.

$$\hat{\Sigma} = \begin{array}{c|ccc} & A_1 & A_2 & A_3 \\ \hline A_1 & 154.67 & 160.00 & 160.00 \\ A_2 & 160.00 & 176.67 & 170.67 \\ A_3 & 160.00 & 170.67 & 170.00 \end{array}$$

On the **main diagonal** of this matrix are the variances within each treatment ($\hat{\theta}^2_{A_i}$). Notice that they are all more or less equal, indicating that we have met the assumption of homogeneity of variance. The **off-diagonal elements** represent the covariances among the treatments (cov_{12}, cov_{13}, and cov_{23}). Notice that these are also more or less equal. (The fact that they are also of the same magnitude as the variances is irrelevant, reflecting merely the very high intercorrelations among treatments.) A pattern of constant variances on the diagonal and constant covariances off the diagonal is referred to as **compound symmetry**. (Again, the relationship between the variances and covariances is irrelevant.) The assumption of compound symmetry of the (*population*) **covariance matrix (Σ)**, of which $\hat{\Sigma}$ is an estimate, represents a sufficient condition underlying a repeated-measures analysis of variance. The more general condition is known as **sphericity**, and you will often see references to that broader assumption. If we have compound symmetry we will meet the sphericity assumption, but it is possible, though not likely in practice, to have sphericity without compound symmetry. (Older textbooks generally make reference to compound symmetry, even though that is too strict an assumption. In recent years the trend has been toward reference to "sphericity," and that is how we will generally refer to it here, though we will return to compound symmetry when we consider mixed models at the end of this chapter.) Without this sphericity assumption, the *F* ratios may not have a distribution given by the distribution of *F* in the tables. Although this assumption applies to any analysis of variance design, when the cells are independent the covariances are always zero, and there is no

problem—we merely need to assume homogeneity of variance. With repeated-measures designs, however, the covariances will not be zero and we need to assume that they are all equal. This has led some people (e.g., Hays, 1981) to omit serious consideration of repeated-measures designs. However, when we do have sphericity, the *F*s are valid; and when we do not, we can use either very good approximation procedures (to be discussed later in this chapter) or alternative methods that do not depend on assumptions about $\sum$. One alternative procedure that does not require any assumptions about the covariance matrix is **multivariate analysis of variance (MANOVA)**. This is a **multivariate procedure**, which is essentially one that deals with multiple dependent variables simultaneously. This procedure, however, requires complete data and is now commonly being replaced by analyses of mixed models, which are introduced in Section 14.12.

Many people have trouble thinking in terms of covariances because they don't have a simple intuitive meaning. There is little to be lost by thinking in terms of correlations. If we truly have homogeneity of variance, compound symmetry reduces to constant correlations between trials.

14.4. ANALYSIS OF VARIANCE APPLIED TO RELAXATION THERAPY

As an example of a simple repeated-measures design, we will consider a study of the effectiveness of relaxation techniques in controlling migraine headaches. The data described here are fictitious, but they are in general agreement with data collected by Blanchard, Theobald, Williamson, Silver, and Brown (1978), who ran a similar, although more complex, study.

In this experiment we have recruited nine migraine sufferers and have asked them to record the frequency and duration of their migraine headaches. After 4 weeks of baseline recording during which no training was given, we had a 6-week period of relaxation training. (Each experimental subject participated in the program at a different time, so such things as changes in climate and holiday events should not systematically influence the data.) For our example we will analyze the data for the last 2 weeks of baseline and the last 3 weeks of training. The dependent variable is the duration (hours/week) of headaches in each of those 5 weeks. The data and the calculations are shown in Table 14.3.[2] It is important to note that I have identified the means with a subscript naming the variable. Thus instead of using the standard "dot notation" (e.g., $\bar{X}_{i\cdot}$ for the Week means), I have used the letter indicating the variable name as the subscript (e.g., the means for Weeks are denoted $\bar{X}_W$ and the means for Subjects are denoted $\bar{X}_S$). As usual, the grand mean is denoted $\bar{X}_{\cdot\cdot}$, and X represents the individual observations.

Table 14.3 Analysis of data on migraine headaches.

(a) Data

	Baseline		**Training**			
Subject	Week 1	Week 2	Week 3	Week 4	Week 5	Subject Means
1	21	22	8	6	6	12.6
2	20	19	10	4	4	11.4
3	17	15	5	4	5	9.2
4	25	30	13	12	17	19.4
5	30	27	13	8	6	16.8
6	19	27	8	7	4	13.0
7	26	16	5	2	5	10.8
8	17	18	8	1	5	9.8
9	26	24	14	8	9	16.2
Week Means	22.333	22.000	9.333	5.778	6.778	13.244

(b) Calculations

$$SS_{total} = \Sigma\left(X - \bar{X}_{..}\right)^2 = (21-13.244)^2 + \ldots + (9-13.244)^2 = 3166.31$$

$$SS_{subjects} = w\Sigma\left(\bar{X}_S - \bar{X}_{..}\right)^2 = 5\left[(12.6-13.244)^2 + \ldots + (16.2-13.244)^2\right] = 486.71$$

$$SS_{weeks} = n\Sigma\left(\bar{X}_W - \bar{X}_{..}\right)^2 = 9\left[(22.333-13.244)^2 + \ldots + (6.778-13.244)^2\right] = 2449.20$$

$$SS_{error} = SS_{total} - SS_{subjects} - SS_{weeks} = 3166.31 - 486.71 - 2449.20 = 230.40$$

(c) Summary table

Source	*df*	SS	MS	*F*
Between subjects	8	486.71		
Within subjects	36	2679.60		85.04*
Weeks	4	2449.20	612.30	
Error	32	230.40	7.20	
Total	44	3166.31		

* $p < .05$

Look first at the data in Table 14.3a. Notice that there is a great deal of variability, but much of that variability comes from the fact that some people have more and/or longer-duration headaches than do others, which really has very little to do with the intervention program. As I have said, what we are able to do with a repeated-measures design but were not able to do with between-subjects designs is to remove this variability from SS_{error}, producing a smaller MS_{error} than we would otherwise have.

From Table 14.3b you can see that SS_{total} is calculated in the usual manner. Similarly, $SS_{subjects}$ and SS_{weeks} are calculated just as main effects always are [take the sum of the squared deviations from the grand mean and multiply by the appropriate constant (i.e., the number of observations

contributing to each mean)]. Finally, the error term is obtained by subtracting $SS_{subjects}$ and SS_{weeks} from SS_{total}.

The summary table is shown in Table 14.3c and looks a bit different from ones you have seen before. In this table I have made a deliberate split into Between-Subject factors and Within-Subject factors. The terms for Weeks and Error are parts of the Within-Subject term, and so are indented under it. (In this design the Between-Subject factor is not further broken down, which is why nothing is indented under it. But wait a few pages and you will see that happen too.) Notice that I have computed an *F* for Weeks but not for subjects, for the reasons given earlier. The *F* value for Weeks is based on 4 and 32 degrees of freedom, and $F_{.05}(4,32) = 2.68$. We can therefore reject H_0: $\mu_1 = \mu_2 = \ldots = \mu_5$ and conclude that the relaxation program led to a reduction in the duration per week of headaches reported by subjects. Examination of the means in Table 14.3 reveals that during the last three weeks of training, the amount of time per week involving headaches was about one-third of what it was during baseline.

You may have noticed that no Subject × Weeks interaction is shown in the summary table. With only one score per cell, the interaction term *is* the error term, and in fact some people prefer to label it *S* × *W* instead of error. To put this differently, in the design discussed here it is impossible to separate error from any possible Subject × Weeks interaction, because they are completely confounded. As we saw in the discussion of structural models, both of these effects, if present, are combined in the expected mean square for error.

I spoke earlier of the assumption of sphericity, or compound symmetry. For the data in the example, the variance–covariance matrix follows, represented by the notation $\hat{\Sigma}$, where the ^ is used to indicate that this is an estimate of the population variance–covariance matrix Σ.

$$\hat{\Sigma} = \begin{matrix} 21.000 & 11.750 & 9.250 & 7.833 & 7.333 \\ 11.750 & 28.500 & 13.750 & 16.375 & 13.375 \\ 9.250 & 13.750 & 11.500 & 8.583 & 8.208 \\ 7.833 & 16.375 & 8.583 & 11.694 & 10.819 \\ 7.333 & 13.375 & 8.208 & 10.819 & 16.945 \end{matrix}$$

Visual inspection of this matrix suggests that the assumption of sphericity is reasonable. The variances on the diagonal range from 11.5 to 28.5, whereas the covariances off the diagonal range from 7.333 to 16.375. Considering that we have only nine subjects, these values represent an acceptable level of constancy. (Keep in mind that the variances do not need to be equal to the covariances; in fact, they seldom are.) A statistical test of this assumption of sphericity was developed by Mauchly (1940) and is given in Winer (1971, p. 596). It would in fact show that we have no basis for rejecting the sphericity hypothesis. Box (1954b), however, showed that regardless of the form of $\sum$, a conservative test on null hypotheses in the repeated-measures analysis of variance is given by comparing F_{obt} against $F_{.05}(1, n-1)$—that is, by acting as though we had only two treatment levels. This test is exceedingly conservative, however, and for most situations you will be better advised to evaluate *F* in the usual way. We will return to this problem later when we consider a much better solution found in Greenhouse and Geisser's (1959) extension of Box's work.

As already mentioned, one of the major advantages of the repeated-measures design is that it allows us to reduce the error term by using the same subject for all treatments. Suppose for a moment that the data illustrated in Table 14.3 had actually been produced by five independent groups of subjects. For such an analysis, SS_{error} would equal 717.11. In this case, we would not be able to pull out a subject term because $SS_{between\,subj}$ would be synonymous with SS_{total}. (A subject total and an individual score are identical.) As a result, differences among subjects would be inseparable from error, and in fact SS_{error} would be the sum of what, for the repeated-measures design, are SS_{error} and $SS_{between\,subj}$ (= 230.4 + 486.71 = 717.11 on 32 + 8 = 40 *df*). This would lead to

$$F = \frac{MS_{weeks}}{MS_{error}} = \frac{612.30}{17.93} = 34.15$$

which, although still significant, is less than one-half of what it was in Table 14.3.

To put it succinctly, subjects differ. When subjects are observed only once, these subject differences contribute to the error term. When subjects are observed repeatedly, we can obtain an estimate of the degree of subject differences and partial these differences out of the error term. In general, the greater the differences among subjects, the higher the correlations between pairs of treatments. The higher the correlations among treatments, the greater the relative power of repeated-measures designs.

We have been speaking of the simple case in which we have one independent variable (other than subjects) and test each subject on every level of that variable. In actual practice, there are many different ways in which we could design a study using repeated measures. For example,

we could set up an experiment using two independent variables and test each subject under all combinations of both variables. Alternatively, each subject might serve under only one level of one of the variables, but under all levels of the other. If we had three variables, the possibilities are even greater. In this chapter we will discuss only a few of the possible designs. If you understand the designs discussed here, you should have no difficulty generalizing to even the most complex problems.

14.5. CONTRASTS AND EFFECT SIZES IN REPEATED MEASURES DESIGNS

As we did in the case of one-way and factorial designs, we need to consider how to run contrasts among means of repeated measures variables. Fortunately there is not really much that is new here. We will again be comparing the mean of a condition or set of conditions against the mean of another condition or set of conditions, and we will be using the same kinds of coefficients that we have used all along.

In our example the first two weeks were Baseline measures, and the last three weeks were Training measures. Our omnibus *F* told us that there were statistically significant differences among the five Weeks, but not where those differences lie. Now I would like to contrast the means of the set of Baseline weeks with the mean of the set of Training weeks. The coefficients that will do this are shown below, along with the means.

	Week 1	Week 2	Week 3	Week 4	Week 5
Coefficient	1/2	1/2	-1/3	-1/3	-1.3
Mean	22.333	22.000	9.333	5.778	6.778

Just as we have been doing, we will define our contrast as

$$\begin{aligned}\hat{\psi} &= \Sigma a_i \bar{X}_i \\ &= \left(\tfrac{1}{2}\right)(22.333) + \left(\tfrac{1}{2}\right)(22.000) + \left(-\tfrac{1}{3}\right)(9.333) + \left(-\tfrac{1}{3}\right)(5.778) + \left(-\tfrac{1}{3}\right)(6.778) \\ &= \frac{22.333 + 22.000}{2} - \frac{9.333 + 5.778 + 6.778}{3} = \frac{44.333}{2} - \frac{21.889}{3} = 22.166 - 7.296 \\ &= 14.870\end{aligned}$$

We can test this contrast with either a *t* or an *F*, but I will use *t* here. (*F* is just the square of *t*.)

$$t = \frac{\hat{\psi}}{\sqrt{\frac{\left(\Sigma a_i^2\right) MS_{error}}{n}}} = \frac{14.870}{\sqrt{\frac{0.833(7.20)}{9}}} = \frac{14.870}{\sqrt{0.667}} = \frac{14.870}{0.816} = 18.21$$

This is a *t* on df_{error} =32 df, and is clearly statistically significant.

Notice that in calculating my *t* I used the MS_{error} from the overall analysis. And this was the same error term that was used to test the Weeks effect. I point that out only because when we come to more complex analyses we will have multiple error terms, and the one to use for a specific contrast is the one that was used to test the main effect of that independent variable.

Effect sizes

Although there was a direct translation from one-way designs to repeated measures designs in terms of testing contrasts among means, the situation is a bit more complicated when it comes to estimating effect sizes. We will continue to define our effect size as

$$\hat{d} = \frac{\hat{\psi}}{s_{error}}$$

There should be no problem with $\hat{\psi}$, because it is the same contrast that we computed above—the difference between the mean of the baseline weeks and the mean of the training weeks. But there are several choices for s_{error}. Kline (2004) gives 3 possible choices for our denominator, but points out that two of these are unsatisfactory either because they ignore the correlation between weeks or because they standardize $\hat{\psi}$ by a standard deviation that is not particularly meaningful. What we will actually do is create an error term that is unique to the particular contrast. We will form a contrast for each subject. That means that for each subject we will calculate the difference between his mean on the baseline weeks and his mean on the training weeks. These are difference scores, which are analogous to the difference scores we computed for a paired sample *t* test. The standard deviation of these difference scores is analogous to the denominator we discussed for computing effect size with paired data when we just had two repeated measures with the *t* test. It is important to note that there is room for argument about the proper term to use to standardize contrasts with repeated measures. See Kline (2004) and Olejnik & Algina (2000).

For our migraine example the first subject would have a difference score of (21 + 22)/2 – (8 + 6 + 6)/3 = 21.5 – 6.667 = 14.833. The complete set of difference scores would be

[14.833, 13.500, 11.333, 13.500, 19.500, 16.667, 17.000, 12.833, 14.667]

The mean of these difference scores is 14.879, which is $\hat{\psi}$. The standard deviation of these difference scores is 2.49. Then our effect size measure is

$$\hat{d} = \frac{\hat{\psi}}{s_{error}} = \frac{14.87}{2.49} = 5.97.$$

This tells us that the severity of headaches during baseline is nearly 6 standard deviations greater than the severity of head aches during training. That is a very large difference, and we can see that just by looking at the data. Remember, in calculating this effect size we have eliminated the variability between participants (subjects) in terms of headache severity. We are in a real sense comparing each individual to him/her self.

14.6. WRITING UP THE RESULTS

In writing up the results of this experiment we could simply say

> To investigate the effects of relaxation therapy on the severity of migraine headaches, 9 participants rated the severity of headaches on each of two weeks before receiving relaxation therapy and for three weeks while receiving therapy. An overall analysis of variance for repeated measures showed a significant difference between weeks ($F(4,32) = 85.04$, $p < .05$). The mean severity rating during baseline weeks was 22.166, which dropped to a mean of 7.296 during training, for a difference of 14.87. A contrast on this difference was significant ($t(32) = 18.21$, $p < .05$). Using the standard deviation of contrast differences for each participant produced an effect size measure of $\boldsymbol{d} = 5.97$, documenting the importance of relaxation therapy in treating migraine headaches.

14.7. ONE BETWEEN-SUBJECTS VARIABLE AND ONE WITHIN-SUBJECTS VARIABLE

Consider the data presented in Table 14.4. These are actual data from a study by King (1986). This study in some ways resembles the one on morphine tolerance by Siegel (1975) that we

examined in Chapter 12. King investigated motor activity in rats following injection of the drug midazolam. The first time that this drug is injected, it typically leads to a distinct decrease in motor activity. Like morphine, however, a tolerance for midazolam develops rapidly. King wished to know whether that acquired tolerance could be explained on the basis of a *conditioned* tolerance related to the physical context in which the drug was administered, as in Siegel's work. He used three groups, collecting the crucial data (presented in Table 14.4) on only the last day, which was the test day. During pretesting, two groups of animals were repeatedly injected with midazolam over several days, whereas the Control group was injected with physiological saline. On the test day, one group—the "Same" group—was injected with midazolam in the *same* environment in which it had earlier been injected. The "Different" group was also injected with midazolam, but in a *different* environment. Finally, the Control group was injected with midazolam for the first time. This Control group should thus show the typical initial response to the drug (decreased ambulatory behavior), whereas the Same group should show the normal tolerance effect—that is, they should decrease their activity little or not at all in response to the drug on the last trial. If King is correct, however, the Different group should respond similarly to the Control group, because although they have had several exposures to the drug, they are receiving it in a novel context and any conditioned tolerance that might have developed will not have the necessary cues required for its elicitation. The dependent variable in Table 14.4 is a measure of ambulatory behavior, in arbitrary units. Again, the first letter of the name of a variable is used as a subscript to indicate what set of means we are referring to.

Table 14.4 Ambulatory behavior by Group and Trial

(a) Data

		Interval						
		1	**2**	**3**	**4**	**5**	**6**	Mean
Control		150	44	71	59	132	74	88.333
		335	270	156	160	118	230	211.500
		149	52	91	115	43	154	100.667
		159	31	127	212	71	224	137.333
		159	0	35	75	71	34	62.333
		292	125	184	246	225	170	207.000
		297	187	66	96	209	74	154.833
		170	37	42	66	114	81	85.000
	Mean	213.875	93.250	96.500	128.625	122.875	130.125	130.875
Same		346	175	177	192	239	140	211.500
		426	329	236	76	102	232	233.500
		359	238	183	123	183	30	186.000
		272	60	82	85	101	98	116.333
		200	271	263	216	241	227	236.333
		366	291	263	144	220	180	244.000
		371	364	270	308	219	267	299.833
		497	402	294	216	284	255	324.667
	Mean	354.625	266.250	221.000	170.000	198.625	178.625	231.521
Different		282	186	225	134	189	169	197.500
		317	31	85	120	131	205	148.167
		362	104	144	114	115	127	161.000
		338	132	91	77	108	169	152.500
		263	94	141	142	120	195	159.167
		138	38	16	95	39	55	63.500
		329	62	62	6	93	67	103.167
		292	139	104	184	193	122	172.333
	Mean	290.125	98.250	108.500	109.000	123.500	138.625	144.667
	Interval mean	286.208	152.583	142.000	135.875	148.333	149.125	169.021

(b) Calculations

$$SS_{total} = \Sigma\left(X - \bar{X}_{...}\right)^2 = (150-169.021)^2 + \ldots + (122-169.021)^2 = 1{,}432{,}292.9$$

$$SS_{subj} = i\Sigma\left(\bar{X}_S - \bar{X}_{...}\right)^2 = 6\left[(88.333-169.021)^2 + \ldots + (172.333-169.021)^2\right] = 670{,}537.1$$

$$SS_{groups} = ni\Sigma\left(\bar{X}_G - \bar{X}_{...}\right)^2 = 8\times 6\left[(130.875-169.021)^2 + \ldots + (144.667-169.021)^2\right] = 285{,}815.0$$

$$SS_{intervals} = ng\Sigma\left(\bar{X}_I - \bar{X}_{...}\right)^2 = 8\times 3\left[(286.208-169.021)^2 + \ldots + (149.125-169.021)^2\right] = 399{,}736.5$$

$$SS_{cells} = n\Sigma\left(\bar{X}_{GI} - \bar{X}_{...}\right)^2 = 8\left[(213.875-169.021)^2 \ldots + (138.625-169.021)^2\right] = 766{,}371.5$$

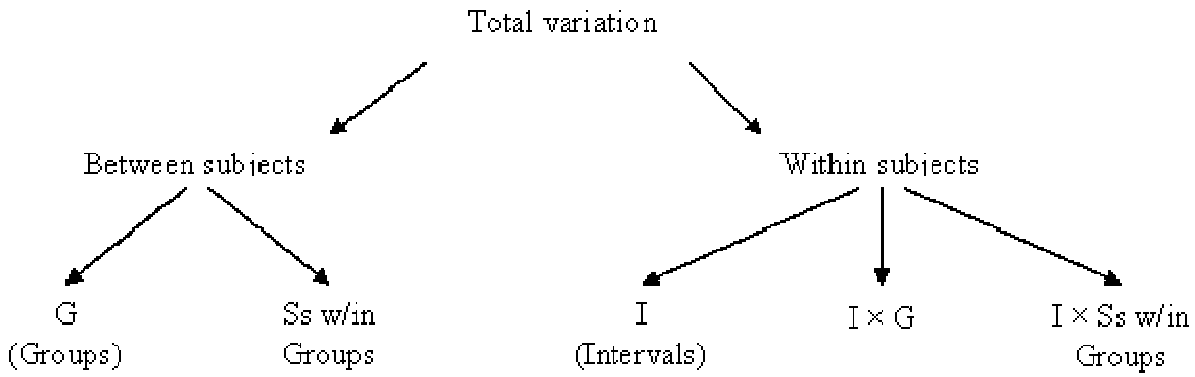
$$SS_{I\times G} = SS_{cells} - SS_{interval} - SS_{groups} = 766{,}371.5 - 285{,}815.0 - 399.736.5 = 80{,}820.0$$

(c) Summary Table

Source	***df***	**SS**	**MS**	***F***
Between subjects	23	670,537.1		
Groups	2	285,815.0	142,907.5	7.80*
*S*s w/in groups**	21	384,722.0	18,320.1	
Within subjects**	120	761,755.8		
Intervals	5	399,736.5	79,947.3	29.85*
$I \times G$	10	80,820.0	8,082.0	3.02*
$I \times S$s w/in groups**	105	281,199.3	2,678.1	
Total	143	1,432,292.9		

* $p < .05$; ** Calculated by subtraction

Because the drug is known to be metabolized over a period of approximately 1 hour, King recorded his data in 5-minute blocks, or Intervals. We would expect to see the effect of the drug increase for the first few intervals and then slowly taper off. Our analysis uses the first six blocks of data. The design of this study can then be represented diagrammatically as



Here we have distinguished those effects that represent differences between subjects from those that represent differences within subjects. When we consider the between-subjects term, we can partition it into differences between groups of subjects (*G*) and differences between subjects in the same group (*S*s w/in groups). The within-subject term can similarly be subdivided into three components—the main effect of Intervals (the repeated measure) and its interactions with the two partitions of the between-subject variation. You will see this partitioning represented in the summary table when we come to it.

Partitioning the between-subjects effects

Let us first consider the partition of the between-subjects term in more detail. From the design of the experiment, we know that this term can be partitioned into two parts. One of these parts is the main effect of Groups (*G*), since the treatments (Control, Same, and Different) involve different groups of subjects. This is not the only source of differences among subjects, however. We have eight different subjects within the control group, and differences among them are certainly between-subjects differences. The same holds for the subjects within the other groups. Here we are speaking of differences among subjects in the same group—that is, *S*s within groups.

If we temporarily ignore intervals entirely (e.g., we simply collect our data over the entire session rather than breaking it down into 5-minute intervals), we can think of the study as producing the following data:

Control	Same	Different
88.333	211.500	197.500
211.500	233.500	148.167
100.667	186.000	161.000
137.333	116.333	152.500
62.333	236.333	159.167
207.000	244.000	63.500
154.833	299.833	103.167
85.000	324.667	172.333
130.875	231.521	144.667

where the “raw scores” in this table are the subject means from Table 14.4. Because each subject is represented only once in these totals, the analysis we will apply here is the same as a one-way analysis of variance on independent groups. Indeed, except for a constant representing the number of scores per subject (which cancels out in the end), the sums of squares for the simple one-way on these data would be the same as those in the actual analysis. The *F* that tests the

main effect of Groups if this were a simple one-way on subject totals would be equal to the one that we will obtain from the full analysis. Thus, the between-subjects partition of the total variation can be seen as essentially a separate analysis of variance, with its own error term (sometimes referred to as $\text{error}_{\text{between}}$) independent of the within-subject effects.

Partitioning the within-subjects effects

Next consider the within-subjects element of the partition of SS_{total}. As we have already seen, this is itself partitioned into three terms. A comparison of the six intervals involves comparisons of scores from the same subject, and thus Intervals is a within-subjects term—it depends on differences within each subject. Since Intervals is a within-subjects term, the interaction of Intervals with Groups is also a within-subjects effect. The third term (Intervals × *S*s within groups) is sometimes referred to as $\text{error}_{\text{within}}$ since it is the error term for the within-subjects effects. The $SS_{\text{Intervals} \times S\text{s w/in groups}}$ term is actually the sum of the sums of squares for the $I \times S$ interactions calculated separately for each group. Thus, it can be seen as logically equivalent to the error term used in the previous design.

The analysis

Before considering the analysis in detail, it is instructive to look at the general pattern of results. Although there are not enough observations in each cell to examine the distributions in any serious way, it is apparent that on any given interval there is substantial variability within groups.

For example, for the second interval in the control group, scores range from 0 to 270. There do not appear to be any extreme outliers, however, as often happens in this kind of research, and the variances within cells, although large, are approximately equal. You can also see that there are large individual differences, with some of the animals consistently showing relatively little ambulatory behavior and some showing a great deal. These are the kinds of differences that will be partialled out by our analysis. Looking at the Interval means, you will see that, as expected, behavior decreased substantially after the first 5-minute interval and then increased slightly during the rest of the session. Finally, looking at the difference between the means for the Control and Same groups, you will see the anticipated tolerance effect, and looking at the Different group, you see that it is much more like the Control group than it is like the Same group. This is the result that King predicted.

Very little needs to be said about the actual calculations in Table 14.4b, since they are really no different from the usual calculations of main and interaction effects. Whether a factor is a between-subjects or within-subjects factor has no bearing on the calculation of its sum of squares, although it does affect its placement in the summary table and the ultimate calculation of the corresponding *F*.

In the summary table in Table 14.4c, the source column reflects the design of the experiment, with SS_{total} first partitioned into $SS_{between\,subj}$ and $SS_{win\,subj}$. Each of these sums of squares is further subdivided. The double asterisks next to the three terms show we calculate these by subtraction ($SS_{win\,subj}$, $SS_{S\times win\,groups}$, and $SS_{I\times S\times win\,groups}$), based on the fact that sums of squares are additive and the whole must be equal to the sum of its parts. This simplifies our work considerably. Thus

$$SS_{w/in\,subj} = SS_{total} - SS_{between\,subj}$$
$$SS_{Ss\,w/in\,groups} = SS_{between\,subj} - SS_{groups}$$
$$SS_{I \times Ss\,w/in\,groups} = SS_{w/in\,subj} - SS_{intervals} - SS_{IG}$$

These last two terms will become error terms for the analysis.

The degrees of freedom are obtained in a relatively straightforward manner. For each of the main effects, the number of degrees of freedom is equal to the number of levels of the variable minus 1. Thus, for Subjects there are 24 - 1 = 23 *df*, for Groups there are 3 - 1 = 2 *df*, and for Intervals there are 6 - 1 = 5 *df*. As for all interactions, the *df* for *I* × *G* is equal to the product of the *df* for the component terms. Thus, $df_{IG} = (6-1)(3-1) = 10$. The easiest way to obtain the remaining degrees of freedom is by subtraction, just as we did with the corresponding sums of squares.

$$df_{w/in\,subj} = df_{total} - df_{between\,subj}$$
$$df_{Ss\,w/in\,groups} = df_{between\,subj} - df_{groups}$$
$$df_{I \times Ss\,w/in\,groups} = df_{w/in\,subj} - df_{intervals} - df_{IG}$$

These *df* can also be obtained directly by considering what these terms represent. Within each subject, we have 6 - 1 = 5 *df*. With 24 subjects, this amounts to $(5)(24) = 120\ df_{w/in\,subj}$. Within each level of the Groups factor, we have 8 - 1 = 7 *df* between subjects, and with three Groups we have $(7)(3) = 21\ df_{w/in\,groups}$. *I* × *S*s w/in groups is really an interaction term, and as such its *df* is simply the product of df_I and $df_{Ss\,w/in\,groups} = (5)(21) = 105$.

Skipping over the mean squares, which are merely the sums of squares divided by their degrees of freedom, we come to *F*. From the column of *F* it is apparent that, as we anticipated, Groups and Intervals are significant. The interaction is also significant, reflecting, in part, the fact that

the Different group was at first intermediate between the Same and the Control group, but that by the second 5-minute interval it had come down to be equal to the Control group. This finding can be explained by a theory of conditioned tolerance. The really interesting finding is that, at least for the later intervals, simply injecting an animal in an environment different from the one in which it had been receiving the drug was sufficient to overcome the tolerance that had developed. These animals respond almost exactly as do animals that had never experienced midazolam. We will return to the comparison of Groups at individual Intervals later.

Assumptions

For the *F* ratios actually to follow the *F* distribution, we must invoke the usual assumptions of normality, homogeneity of variance, and sphericity of $\hat{\Sigma}$. For the *between-subjects* term(s), this means that we must assume that the variance of subject means within any one level of Group is the same as the variance of subject means within every other level of Group. If necessary, this assumption can be tested by calculating each of the variances and testing using either F_{max} on $(g, n-1)df$ or, preferably, the test proposed by Levene (1960) or O'Brien (1981), which were referred to in Chapter 7. In practice, however, the analysis of variance is relatively robust against reasonable violations of this assumption (see Collier, Baker, and Mandeville, 1967; and Collier, Baker, Mandeville, and Hayes, 1967). Because the groups are independent, compound symmetry, and thus sphericity, of the covariance matrix is assured if we have homogeneity of variance, since all off-diagonal entries will be zero.

For the *within-subjects* terms we must also consider the usual assumptions of homogeneity of variance and normality. The homogeneity of variance assumption in this case is that the $I \times S$ interactions are constant across the Groups, and here again this can be tested using $F_{\max}$ on g and $(n-1)(i-1)\,df$. (You would simply calculate an $I \times S$ interaction for each group—equivalent to the error term in Table 14.3—and test the largest against the smallest.) For the within-subjects effects, we must also make assumptions concerning the covariance matrix.

There are two assumptions on the covariance matrix (or matrices). Again, we will let $\hat{\Sigma}$ represent the matrix of variances and covariances among the levels of *I* (Intervals). Thus with six intervals,

$$\hat{\Sigma} = \begin{array}{c|cccccc} & \boldsymbol{I_1} & \boldsymbol{I_2} & \boldsymbol{I_3} & \boldsymbol{I_4} & \boldsymbol{I_5} & \boldsymbol{I6} \\ \hline & \hat{\sigma}_{11} & \hat{\sigma}_{12} & \hat{\sigma}_{13} & \hat{\sigma}_{14} & \hat{\sigma}_{15} & \hat{\sigma}_{16} \\ & \hat{\sigma}_{21} & \hat{\sigma}_{22} & \hat{\sigma}_{23} & \hat{\sigma}_{24} & \hat{\sigma}_{25} & \hat{\sigma}_{26} \\ & \hat{\sigma}_{31} & \hat{\sigma}_{32} & \hat{\sigma}_{33} & \hat{\sigma}_{34} & \hat{\sigma}_{35} & \hat{\sigma}_{36} \\ & \hat{\sigma}_{41} & \hat{\sigma}_{42} & \hat{\sigma}_{43} & \hat{\sigma}_{44} & \hat{\sigma}_{45} & \hat{\sigma}_{46} \\ & \hat{\sigma}_{51} & \hat{\sigma}_{52} & \hat{\sigma}_{53} & \hat{\sigma}_{54} & \hat{\sigma}_{55} & \hat{\sigma}_{56} \\ & \hat{\sigma}_{61} & \hat{\sigma}_{62} & \hat{\sigma}_{63} & \hat{\sigma}_{64} & \hat{\sigma}_{65} & \hat{\sigma}_{66} \end{array}$$

For each Group we would have a separate population variance–covariance matrix Σ_{G_i}. (Σ and Σ_{G_i} are estimated by $\hat{\Sigma}$ and $\hat{\Sigma}_{G_i}$, respectively.) For $MS_{I \times S \text{ within groups}}$ to be an appropriate error term, we will first assume that the individual variance–covariance matrices (Σ_{G_i}) are the same

for all levels of *G*. This can be thought of as an extension (to covariances) of the common assumption of homogeneity of variance.

The second assumption concerning covariances deals with the overall matrix Σ, where Σ is the pooled average of the Σ_G. (For equal sample sizes in each group, an entry in Σ will be the average of the corresponding entries in the individual Σ_G matrices.) A common and sufficient, but not necessary, assumption is that the matrix exhibits compound symmetry—meaning, as I said earlier, that all the variances on the main diagonal are equal, and all the covariances off the main diagonal are equal. Again, the variances do not have to equal the covariances, and usually will not. This assumption is in fact more stringent than necessary. All that we really need to assume is that the standard errors of the differences between pairs of Interval means are constant—in other words, that $\sigma^2_{I_i - I_j}$ is constant for all *i* and *j* ($j \neq i$). This sphericity requirement is met automatically if Σ exhibits compound symmetry, but other patterns of Σ will also have this property. For a more extensive discussion of the covariance assumptions, see Huynh and Feldt (1970) and Huynh and Mandeville (1979); a particularly good discussion can be found in Edwards (1985, pp. 327–329, 336–339).

Adjusting the degrees of freedom

Box (1954a) and Greenhouse and Geisser (1959) considered the effects of departure from this sphericity assumption on $\sum$. They showed that regardless of the form of $\sum$, the *F* ratio from the within-subjects portion of the analysis of variance will be approximately distributed as *F* on

$(i - 1)\varepsilon, g(n - 1)(i - 1)\varepsilon$

df for the Interval effect and

$(g - 1)(i - 1)\varepsilon, g(n - 1)(i - 1)\varepsilon$

df for the $I \times G$ interaction, where i = the number of intervals and ε is estimated by

$$\hat{\varepsilon} = \frac{i^2\left(\bar{s}_{jj} - \bar{s}\right)^2}{(i-1)\left(\sum s_{jk}^2 - 2i\sum \bar{s}_j^2 + i^2\bar{s}^2\right)}$$

Here,

$\bar{s}_{jj}$ = the mean of the entries on the main diagonal of $\hat{\Sigma}$

$\bar{s}$ = the mean of all entries in $\hat{\Sigma}$

s_{jk} = the jkth entry in $\hat{\Sigma}$

$\bar{s}_j$ = the mean of all entries in the jth row of $\hat{\Sigma}$

The effect of using $\hat{\varepsilon}$ is to decrease both df_{effect} and df_{error} from what they would normally be. Thus $\hat{\varepsilon}$ is simply the proportion by which we reduce them. Greenhouse and Geisser recommended that we adjust our degrees of freedom using $\hat{\varepsilon}$. They further showed that when the sphericity assumptions are met, $\varepsilon = 1$, and as we depart more and more from sphericity, ε approaches $1/(i - 1)$ as a minimum.

There is some suggestion that for large values of ε, even using $\hat{\varepsilon}$ to adjust the degrees of freedom can lead to a conservative test. Huynh and Feldt (1976) investigated this correction and recommended a modification of $\hat{\varepsilon}$ when there is reason to believe that the true value of ε lies near or above 0.75. Huynh and Feldt, as later corrected by Lecoutre (1991) defined

$$\tilde{\varepsilon} = \frac{(N - g + 1)(i - 1)\hat{\varepsilon} - 2}{(i - 1)\left[N - g - (i - 1)\hat{\varepsilon}\right]}$$

where $N = n \times g$. (Chen and Dunlap (1994) later confirmed Lecoutre's correction to the original Huynh and Feldt formula.[3])We then use $\hat{\varepsilon}$ or $\tilde{\varepsilon}$, depending on our estimate of the true value of ε. (Under certain circumstances, $\tilde{\varepsilon}$ will exceed 1, at which point it is set to 1.)

A test on the assumption of sphericity has been developed by Mauchly (1940) and evaluated by Huynh and Mandeville (1979) and by Keselman, Rogan, Mendoza, and Breen (1980), who point to its extreme lack of robustness. This test is available on SPSS, SAS, and other software, and is routinely printed out. Because tests of sphericity are likely to have serious problems when we need them the most, it has been suggested that we *always* use the correction to our degrees of freedom afforded by $\hat{\varepsilon}$ or $\tilde{\varepsilon}$, whichever is appropriate, or use a multivariate procedure to be discussed later. This is a reasonable suggestion and one worth adopting.

For our data, the F value for Intervals ($F = 29.85$) is such that its interpretation would be the same regardless of the value of ε, since the Interval effect will be significant even for the lowest possible *df*. If the assumption of sphericity is found to be invalid, however, alternative treatments would lead to different conclusions with respect to the $I \times G$ interaction. For King's data, the Mauchly's sphericity test, as found from SPSS, indicates that the assumption has been violated, and therefore it is necessary to deal with the problem resulting from this violation.

We can calculate $\hat{\varepsilon}$ and $\tilde{\varepsilon}$ and evaluate F on the appropriate *df*. The pooled variance–covariance matrix (averaged across the separate matrices) is presented in Table 14.5. (I have not presented

the variance–covariance matrices for the several groups because they are roughly equivalent and because each of the elements of the matrix is based on only eight observations.)

From Table 14.5 we can see that our values of $\hat{\varepsilon}$ and $\tilde{\varepsilon}$ are .6569 and .7508, respectively. Since these are in the neighborhood of .75, we will follow Huynh and Feldt's suggestion and use $\tilde{\varepsilon}$. In this case, the degrees of freedom for the interaction are

$$(g - 1)(i - 1)(.7508) = 7.508$$

and

$$g(n - 1)(i - 1)(.7508) = 78.834$$

The exact critical value of $F_{.05}(7.508, 78.834)$ is 2.09, which means that we will reject the null hypothesis for the interaction. Thus, regardless of any problems with sphericity, all the effects in this analysis are significant. (They would also be significant if we used $\hat{\varepsilon}$ instead of $\tilde{\varepsilon}$.)

Table 14.5 Variance-covariance matrix and calculation of $\hat{\varepsilon}$ and $\tilde{\varepsilon}$

	Interval						
	1	2	3	4	5	6	Mean
	6388.173	4696.226	2240.143	681.649	2017.726	1924.066	2991.330
	4696.226	7863.644	4181.476	2461.702	2891.524	3531.869	4271.074
	2240.143	4181.476	3912.380	2696.690	2161.690	3297.762	3081.690
	681.649	2461.702	2696.690	4601.327	2248.600	3084.589	2629.093
	2017.726	2891.524	2161.690	2248.600	3717.369	989.310	2337.703
	1924.066	3531.869	3297.762	3084.589	989.310	5227.649	3009.208

$$\bar{s}_{jk} = \frac{6388.173 + 7863.644 \ldots + 5227.649}{6} = 5285.090$$

$$\bar{s} = \frac{6388.173 + 4696.226 + \ldots + 989.310 + 5227.649}{36} = 3053.350$$

$$\Sigma s_{jk}^2 = 6388.173^2 + 4696.226^2 + \ldots + 5227.649^2 = 416,392,330$$

$$\Sigma \bar{s}_j^2 = 2991.330^2 + \ldots + 3009.208^2 = 58,119,260$$

$$\hat{\varepsilon} = \frac{i^2 \left(\bar{s}_{jk} - \bar{s}\right)^2}{(i-1)\left(\Sigma s_{jk}^2 - 2i\Sigma \bar{s}_j^2 + i^2 \bar{s}^2\right)}$$

$$= \frac{36(5285.090 - 3053.350)^2}{(6-1)\left[416,392,330 - (2)(6)(58,119,260) + (36)\left(3053.350^2\right)\right]}$$

$$= \frac{179,303,883}{5[416,392,330 - 697,431,120 + 335,626,064]} = 0.6569$$

$$\tilde{\varepsilon} = \frac{(N - g + 1)(i-1)\hat{\varepsilon} - 2}{(i-1)\left[N - g - (i-1)\hat{\varepsilon}\right]}$$

$$= \frac{(24 - 3 + 1)(5)(0.6569) - 2}{5\left[24 - 3 - 5(0.6569)\right]} = \frac{70.259}{93.5775} = 0.7508$$

Simple effects

The Interval × Group interaction is plotted in Figure 14.2; the interpretation of the data is relatively clear. It is apparent that the Same group consistently performs above the level of the other two groups—that is, the conditioned tolerance to midazolam leads to greater activity in that group than in the other groups. It is also clear that activity decreases noticeably after the first 5-minute interval (during which the drug is having its greatest effect). The interaction appears to be produced by the fact that the Different group is intermediate between the other two groups during the first interval, but it is virtually indistinguishable from the Control group thereafter. In addition, the Same group continues declining until at least the fourth interval, whereas the other

two groups drop precipitously and then level off. Simple effects will prove useful in interpreting these results, especially in terms of examining group differences during the first and the last intervals. Simple effects will also be used to test for differences between intervals within the Control group, but only for purposes of illustration—it should be clear that Interval differences exist within each group.

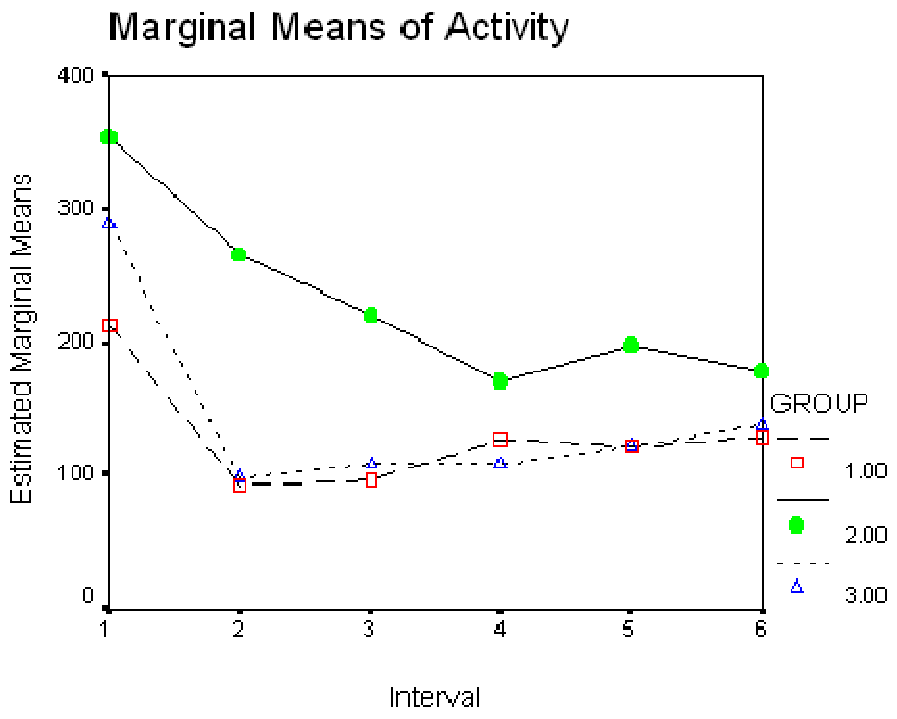


Figure 14.2 Interval × Group interaction for data from Table 14.4

As I have suggested earlier, the Greenhouse and Geisser and the Huynh and Feldt adjustments to degrees of freedom appear to do an adequate job of correcting for problems with the sphericity assumption when testing for overall main effects or interactions. However, a serious question about the adequacy of the adjustment arises when we consider within-subjects simple effects (Boik, 1981; Harris, 1985). The traditional approach to testing simple effects (see Howell, 1987)

involves testing individual within-subjects contrasts against a pooled error term ($MS_{I \times S \text{ w/in groups}}$). If there are problems with the underlying assumption, this error term will sometimes underestimate and sometimes overestimate what would be the proper denominator for *F*, playing havoc with the probability of a Type I error. For that reason we are going to adopt a different, and in some ways simpler, approach.

The approach we will take follows the advice of Boik that a separate error term be derived for each tested effect. Thus, when we look at the simple effect of Intervals for the Control condition, for example, the error term will speak specifically to that effect and will not pool other error terms that apply to other simple effects. In other words, it will be based solely on the Control group. We can test the Interval simple effects quite easily by running separate repeated-measures analyses of variance for each of the groups. For example, we can run a one-way repeated-measures analysis on Intervals for the Control group, as discussed in Section 14.4. We can then turn around and perform similar analyses on Intervals for the Same and Different groups separately. These results are shown in Table 14.6. In each case the Interval differences are significant, even after we correct the degrees of freedom using $\hat{\varepsilon}$ or $\tilde{\varepsilon}$, whichever is appropriate.

Table 14.6 Calculation of within-subjects simple effects for data from King (1986)

(a) Interval at Control

Source	***df***	**SS**	**MS**	***F***
Between subjects	7	134,615.58		
Interval	5	76,447.25	15,289.45	5.69*
Error	35	93,998.42	2685.67	
Total	47	305,061.25		

*$p < .05$; $\hat{\varepsilon} = .404$; $\tilde{\varepsilon} = .570$

(b) Interval at Same

Source	*df*	SS	MS	*F*
Between subjects	7	175,600.15		
Interval	5	193,090.85	38,618.17	11.10*
Error	35	121,714.98	3477.57	
Total	47	490,405.98		

*$p < .05$; $\hat{\varepsilon} = .578$; $\tilde{\varepsilon} = 1.00$

(c) Interval at Different

Source	*df*	SS	MS	*F*
Between subjects	7	74,506.33		
Interval	5	211,018.42	42,203.68	22.56*
Error	35	65,485.92	1871.03	
Total	47	351,010.67		

*$p < .05$; $\hat{\varepsilon} = .598$; $\tilde{\varepsilon} = 1.00$

If you look at the within-subject analyses in Table 14.6, you will see that the average MS_{error} is (2685.669 + 3477.571 + 1871.026)/3 = 2678.089, which is $MS_{I \times S \text{ within groups}}$ from the overall analysis found on page xxx. Here these denominators for the *F* ratios are noticeably different from what they would have been had we used the pooled term, which is the traditional approach. You can also verify with a little work that the $MS_{Interval}$ terms for each analysis are the same as those that we would compute if we followed the usual procedures for obtaining simple effects mean squares.

For the between-subjects simple effects (e.g., Groups at Interval 1) the procedure is more complicated. Although we could follow the within-subject example and perform separate

analyses at each Interval, we would lose considerable degrees of freedom unnecessarily. Here it is usually legitimate to pool error terms, and it is generally wise to do so.

For this example we will examine the simple effects of Group at Interval 1 and Group at Interval 6. The original data can be found in Table 14.4 on page xxx. The sums of squares for these effects are

$$SS_{G \text{ at Int. } 1} = 8\left[(213.875-286.208)^2+(354.625-286.208)^2+(290.125-286.208)^2\right]=79,426.33$$
$$SS_{G \text{ at Int. } 6} = 8\left[(130.125-149.125)^2+(178.625-149.125)^2+(138.625-149.125)^2\right]=10,732.00$$

Testing the simple effects of between-subjects terms is a little trickier. Consider for a moment the simple effect of Group at Interval 1. This is essentially a one-way analysis of variance with no repeated measures, since the Group means now represent the average of single—rather than repeated—observations on subjects. Thus, subject differences are confounded with experimental error. In this case, the appropriate error sum of squares is $SS_{\text{w/in cell}}$, where, from Table 14.4,

$$\begin{aligned} SS_{\text{w/in cell}} &= SS_{\text{Ss w/in group}} + SS_{I\times \text{Ss w/in groups}} \\ &= 384,722.03 + 281,199.34 = 665,921.37 \end{aligned}$$

and

$$MS_{\text{w/in cell}} = \frac{SS_{\text{w/in cell}}}{df_{\text{Ss w/in group}} + df_{I\times \text{Ss w/in group}}}$$

$$= \frac{665,921.37}{21+105} = 5285.09$$

It may be easier for you to understand why we need this special $MS_{w/in cell}$ error term if you think about what it really represents. If you were presented with only the data for Interval 1 in Table 14.4 and wished to test the differences among the three groups, you would run a standard one-way analysis of variance, and the MS_{error} would be the average of the variances within each of the three groups. Similarly, if you had only the data from Interval 2, Interval 3, and so on, you would again average the variances within the three treatment groups. The $MS_{w/in cell}$ that we have just finished calculating is in reality the average of the error terms for these six different sets (Intervals) of data. As such, it is the average of the variance within each of the 18 cells.

We can now proceed to form our *F* ratios.

$$F_{Grp.\,Int.\,1} = \frac{MS_{Grp.\,Int.\,1}}{MS_{w/in cell}} = \frac{79,426.33/2}{5285.09} = 7.51$$

$$F_{Grp.\,Int.\,6} = \frac{MS_{Grp.\,Int.\,6}}{MS_{w/in cell}} = \frac{10,732/2}{5285.09} = 1.02$$

A further difficulty arises in the evaluation of *F*. Since $MS_{w/in cell}$ also represents the sum of two *heterogeneous* sources of error [as can be seen by examination of the *E*(MS) for *S*s w/in groups and *I* × *S*s w/in groups], our *F* will not be distributed on 2 and 126 *df*. We will get ourselves out of this difficulty in the same way we did when we faced a similar problem concerning *t* in Chapter 7. We will simply calculate the relevant *df* against which to evaluate *F*—more precisely, we will calculate a statistic denoted as f' and evaluate F_{obt} against $F_{.05}(a-1, f')$. In this case, the value of f' is given by Welch (1938) and Satterthwaite (1946) as

$$f' = \frac{(u+v)^2}{\frac{u^2}{df_u} + \frac{v^2}{df_v}}$$

where

$$u = SS_{\text{Ss w/in groups}}$$
$$v = SS_{\text{Ls \& Ss w/in groups}}$$

and df_u and df_v are the corresponding degrees of freedom. For our example,

$$u = 384,722.03 \quad df_u = 21$$
$$v = 281,199.34 \quad df_v = 105$$
$$f' = \frac{(384,722.03 + 281,199.34)^2}{\frac{384,722.03^2}{21} + \frac{281,199.34^2}{105}} = 56.84$$

Rounding to the nearest integer gives f' = 57. Thus, our *F* is distributed on (*g* - 1, f') = (2, 57) *df* under H_0. For 2 and 57 *df*, $F_{.05}$ = 3.16. Only the difference at Interval 1 is significant. By the end of 30 minutes, the three groups were performing at equivalent levels. It is logical to conclude that somewhere between the first and the sixth interval the three groups become nonsignificantly different, and many people test at each interval to find that point. However, I strongly recommend against this practice as a general rule. We have already run a number of significance tests, and running more of them serves only to increase the error rate. Unless there is an important theoretical reason to determine the point at which the group differences become nonsignificant—and I suspect that there are very few such cases—then there is nothing to be gained by testing each interval. Tests should be carried out to answer important questions, not to address idle curiosity or to make the analysis look "complete."

Multiple comparisons

Several studies have investigated the robustness of multiple-comparison procedures for testing differences among means on the within-subjects variable. Maxwell (1980) studied a simple repeated-measures design with no between-subject component and advised adopting multiple-comparison procedures that do not use a pooled error term. We discussed such a procedure (the Games–Howell procedure) in Chapter 12. (I did use a pooled error term in the analysis of the migraine study, but there it was reasonable to assume homogeneity of variance and I was using all of the weeks. If I had only been running a contrast involving three of the weeks, I would seriously consider calculating an error term based on just the data from those weeks.)

Keselman and Keselman (1988) extended Maxwell's work to designs having one between-subject component and made a similar recommendation. In fact, they showed that when the Groups are of different sizes and sphericity is violated, familywise error rates can become very badly distorted. In the simple effects procedures that we have just considered, I recommended using separate error terms by running one-way repeated-measures analyses for each of the groups. For subsequent multiple-comparison procedures exploring those simple effects, especially with unequal sample sizes, it would probably be wise to employ the Games–Howell procedure using those separate covariance matrices. In other words, to compare Intervals 3 and 4 for the Control group, you would generate your error term using only the Intervals 3 and 4 data from just the Control group.

Myers (1979) has suggested making post hoc tests on a repeated measure using paired *t*-tests and a Bonferroni correction. (This is essentially what I did for the migraine example, though a Bonferroni correction was not necessary because I ran only one contrast.) Maxwell (1980)

showed that this approach does a good job of controlling the familywise error rate, and Baker and Lew (1987) showed that it generally compared well against Tukey's test in terms of power. Baker proposed a simple modification of the Bonferroni (roughly in line with that of Holm) that had even greater power.

14.8. TWO BETWEEN-SUBJECTS VARIABLES AND ONE WITHIN-SUBJECTS VARIABLE

The basic theory of repeated-measures analysis of variance has already been described in the discussion of the previous designs. However, experimenters commonly plan experiments with three or more variables, some or all of which represent repeated measures on the same subjects. We will briefly discuss the analysis of these designs. The calculations are straight forward, because the sums of squares for main effects and interactions are obtained in the usual way and the error terms are obtained by subtraction.

We will not consider the theory behind these designs at any length. Essentially, it amounts to the extrapolation of what has already been said about the two-variable case. For an excellent discussion of the underlying statistical theory see Winer (1971) or Maxwell and Delaney (2004).

I will take as an example a study by St. Lawrence, Brasfield, Shirley, Jefferson, Alleyne, and O'Bannon (1995) on an intervention program to reduce the risk of HIV infection among African-American adolescents. The study involved a comparison of two approaches, one of which was a

standard 2-hour educational program used as a control condition (EC) and the other was an 8-week behavioral skills training program (BST). Subjects were Male and Female adolescents, and measures were taken at Pretest, Posttest, and 6 and 12 months follow-up (FU6 and FU12). There were multiple dependent variables in the study, but the one that we will consider is log(freq + 1), where freq is the frequency of condom-protected intercourse[4]. This is a 2 × 2 × 4 repeated-measures design, with Intervention and Sex as between-subjects factors and Time as the within-subjects factor. This design may be diagrammed as follows, where G_i represents the *i*th group of subjects.

	Behavioral Skills Training				**Educational Control**			
	Pretest	**Posttest**	**FU6**	**FU12**	**Pretest**	**Posttest**	**FU6**	**FU12**
Male	G_1	G_1	G_1	G_1	G_2	G_2	G_2	G_2
Female	G_3	G_3	G_3	G_3	G_4	G_4	G_4	G_4

The raw data and the necessary summary tables of cell totals are presented in Table 14.7a. (These data have been generated to closely mimic the data reported by St. Lawrence et al., though they had many more subjects. Decimal points have been omitted.) In Table 14.7b are the calculations for the main effects and interactions. Here, as elsewhere, the calculations are carried out exactly as they are for any main effects and interactions.

Table 14.7 Data and analysis of study by St. Lawrence et al. (1995)

(a) Data

	Male				Female			
	Pretest	Posttest	FU6	FU12	Pretest	Posttest	FU6	FU12
Behavioral Skill Training	7	22	13	14	0	6	22	26
	25	10	17	24	0	16	12	15
	50	36	49	23	0	8	0	0
	16	38	34	24	15	14	22	8
	33	25	24	25	27	18	24	37
	10	7	23	26	0	0	0	0
	13	33	27	24	4	27	21	3
	22	20	21	11	26	9	9	12
	4	0	12	0	0	0	14	1
	17	16	20	10	0	0	12	0
Educational Control	0	0	0	0	15	28	26	15
	69	56	14	36	0	0	0	0
	5	0	0	5	6	0	23	0
	4	24	0	0	0	0	0	0
	35	8	0	0	25	28	0	16
	7	0	9	37	36	22	14	48
	51	53	8	26	19	22	29	2
	25	0	0	15	0	0	5	14
	59	45	11	16	0	0	0	0
	40	2	33	16	0	0	0	0

Group × Sex × Time means

		Pretest	Posttest	FU6	FU12	Mean
BST	Male	19.7	20.7	24.0	18.1	20.625
BST	Female	7.2	9.8	13.6	10.2	10.200
EC	Male	29.5	18.8	7.5	15.1	17.725
EC	Female	10.1	10.0	9.7	9.5	9.825
Mean		16.625	14.825	13.700	13.225	14.594

Group × Sex means

	Male	Female	Mean
BST	20.625	10.200	15.412
EC	17.725	9.825	13.775
Mean	19.175	10.012	14.594

(b) Calculations

$$SS_{total} = \Sigma(X - \bar{X})^2 = (7-14.594)^2 + \ldots + (0-14.594)^2 = 35404.594$$

$$SS_{subj} = t\Sigma(\bar{X}_{Subj} - \bar{X})^2 = 4\left[(14-14.594)^2 + \ldots + (0-14.594)^2\right] = 21490.344$$

$$SS_{group} = ntsg\Sigma\left(\bar{X}_G - \bar{X}\right)^2 = 10\times4\times2\left[(15.412-14.594)^2 + (13.775-14.594)^2\right] = 107.256$$

$$SS_{Sex} = ntg\Sigma\left(\bar{X}_{Sex} - \bar{X}\right)^2 = 10\times4\times2\left[(19.175-14.594)^2 + (10.012-14.594)^2\right] = 3358.056$$

$$SS_{cells\ GS} = nt\Sigma(\bar{X}_{cells\ GS} - \bar{X})^2 = 10\times4\left[(20.625-14.594)^2 + \ldots + (9.825-14.594)^2\right] = 3529.069$$

$$SS_{GS} = SS_{cells\ GS} - SS_G - SS_S = 3529.069 - 107.256 - 3358.056 = 63.757$$

$$SS_{time} = ngs\Sigma\left(\bar{X}_T - \bar{X}\right)^2 = 10\times2\times2\left[(16.625-14.594)^2 + \ldots + (13.225-14.594)^2\right] = 274.069$$

$$SS_{cells\ TG} = ns\Sigma(\bar{X}_{cells\ TG} - \bar{X})^2 = 10\times2\left[(13.45-14.594)^2 + \ldots + (12.300-14.594)^2\right] = 1759.144$$

$$SS_{TG} = SS_{cells\ TG} - SS_T - SS_G = 1759.144 - 274.069 - 107.256 = 1377.819$$

$$SS_{cells\ TS} = ng\Sigma(\bar{X}_{cells\ TS} - \bar{X})^2 = 10\times2\left[(24.60-14.594)^2 + \ldots + (9.85-14.594)^2\right] = 4412.044$$

$$SS_{TS} = SS_{cells\ TS} - SS_T - SS_S = 4412.044 - 274.069 - 3358.056 = 779.919$$

$$SS_{cells\ GTS} = n\Sigma(\bar{X}_{cells\ GTS} - \bar{X})^2 = 10\left[(19.7-14.594)^2 + \ldots + (9.50-14.594)^2\right] = 6437.294$$

$$SS_{GTS} = SS_{cells\ GTS} - SS_G - SS_T - SS_S - SS_{GT} - SS_{GS} - SS_{TS}$$
$$= 6437.294 - 107.256 - 274.069 - 3358.056 - 1377.819 - 63.757 - 779.919 = 476.419$$

(c) Summary Table

Source	df	SS	MS	*F*
Between subjects	39	21,490.344		
Group (Condition)	1	107.256	107.256	0.21
Sex	1	3358.056	3358.056	6.73*
$G \times S$	1	63.757	63.757	0.13
Ss w/in groups**	36	17,961.275	498.924	
Within subjects**	120	13,914.250		
Time	3	274.069	91.356	0.90
$T \times G$	3	1377.819	459.273	4.51*
$T \times S$	3	779.919	259.973	2.55
$T \times G \times S$	3	476.419	158.806	1.56

$T \times S$s w/in groups**	108	11,006.025	101.908
Total	159	35,404.594	

*$p < .05$ ** Obtained by subtraction

The summary table for the analysis of variance is presented in Table 14.7c. In this table, the ** indicate terms that were obtained by subtraction. Specifically,

$$SS_{w/in\ subj} = SS_{total} - SS_{between\ subj}$$
$$SS_{Ss\ w/in\ groups} = SS_{between\ subj} - SS_{G} - SS_{S} - SS_{GS}$$
$$SS_{T \times Ss\ w/in\ groups} = SS_{w/in\ subj} - SS_{T} - SS_{TG} - SS_{TS} - SS_{TGS}$$

These last two terms are the error terms for between-subjects and within-subjects effects, respectively. That these error terms are appropriate is shown by examining the expected mean squares presented in Table 14.8 on page xxx[5]. For the expected mean squares of random and mixed models, see Kirk (1968) or Winer (1971).

From the column of *F* in the summary table of Table 14.7c, we see that the main effect of Sex is significant, as is the Time × Group interaction. Both of these results are meaningful. As you will recall, the dependent variable is a measure of the frequency of use of condoms (log(freq + 1)). Examination of the means reveals adolescent girls report a lower frequency of use than adolescent boys. That could mean either that they have a lower frequency of intercourse, or that they use condoms a lower percentage of the time. Supplementary data supplied by St. Lawrence et al. show that females do report using condoms a lower percentage of the time than males, but not enough to account for the difference that we see here. Apparently what we are seeing is a reflection of the reported frequency of intercourse.

The most important result in this summary table is the Time × Group interaction. This is precisely what we would be looking for. We don't really care about a Group effect, because we

would like the groups to be equal at pretest, and that equality would dilute any overall group difference. Nor do we particularly care about a main effect of Time, because we expect the Control group not to show appreciable change over time, and that would dilute any Time effect. What we really want to see is that the BST group increases their use over time, whereas the EC group remains constant. That is an interaction, and that is what we found.

Table 14.8 Expected mean squares with *A*, *B*, and *C* fixed

Source	df	SS
Between subjects	*abn*-1	
A	*a*-1	$\sigma_e^2 + c\sigma_s^2 + nbc\theta_\alpha^2$
B	*b*-1	$\sigma_e^2 + c\sigma_s^2 + nac\theta_\beta^2$
AB	*(a-1)(b*-1)	$\sigma_e^2 + c\sigma_s^2 + nc\theta_{\alpha\beta}^2$
Ss w/in groups	*ab*(*n*-1)	$\sigma_e^2 + c\sigma_s^2$
Within subjects	*abn*(*c*-1)	
C	*c*-1	$\sigma_e^2 + \sigma_{ys}^2 + nab\theta_\gamma^2$
AC	(*a*-1)(*c*-1)	$\sigma_e^2 + \sigma_{ys}^2 + nb\theta_{\alpha\gamma}^2$
BC	(*b*-1)(*c*-1)	$\sigma_e^2 + \sigma_{ys}^2 + na\theta_{\beta\gamma}^2$
ABC	(*a*-1) (*b*-1)(*c*-1)	$\sigma_e^2 + \sigma_{ys}^2 + n\theta_{\alpha\beta\gamma}^2$
C × Ss w/in groups	*ab*(*n*-1)(*c*-1)	$\sigma_e^2 + \sigma_{ys}^2$
Total	*N*-1	

Simple effects for complex repeated-measures designs

In the previous example we saw that tests on within-subjects effects were occasionally disrupted by violations of the sphericity assumption, and we took steps to work around this problem. We will have much the same problem with this example.

The cell means plotted in Figure 14.3 reveal the way in which frequency of condom use changes over time for the two treatment conditions and for males and females separately. It is clear from this figure that the data do not tell a simple story.

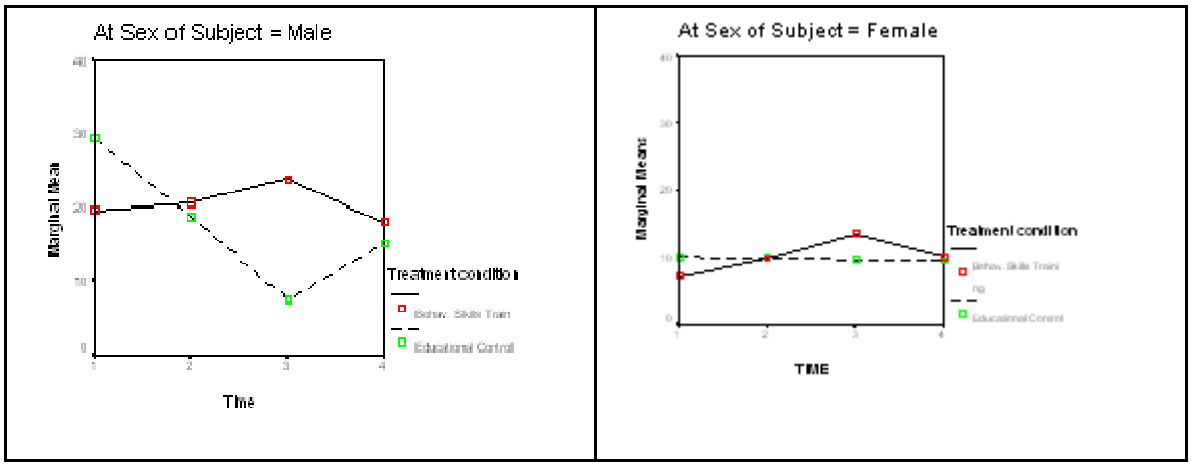


Figure 14.3 Frequency of condom use as a function of Sex and Condition

We are again going to have to distinguish between simple effects on between-subject factors and simple effects on within-subject factors. We will start with between-subject simple effects. We have three different between-subjects simple effects that we could examine—namely; the simple main effects of Condition and Sex at each Time, and the Sex × Condition simple interaction

effect at each Time. For example, we might wish to check that the two Conditions (BST and EC) do not differ at pretest. Again, we might also want to test that they do differ at FU6 and/or at FU12. Here we are really dissecting the Condition × Time interaction effect, which we know from Table 14.7 to be significant.

By far the easiest way to test these between-subjects effects is to run separate two-way (Condition × Sex) analyses at each level of the Time variable. These four analyses will give you all three simple effects at each Time with only minor effort. You can then accept the *F* values from these analyses, as I have done here for convenience, or you can pool the error terms from the four separate analyses and use that pooled error term in testing the mean square for the relevant effect. If these terms are heterogeneous, you would be wise not to pool them. On the other hand, if they represent homogeneous sources of variance, they may be pooled, giving you more degrees of freedom for error. For these effects you don't need to worry about sphericity because each simple effect is calculated on only one level of the repeated-measures variable.

The within-subjects simple effects are handled in much the same way. For example, there is some reason to look at the simple effects of Time for each Condition separately to see whether the EC condition shows changes over time in the absence of a complete intervention. Similarly, we would like to see how the BST condition changes with time. However, we want to include Sex as an effect in both of these analyses so as not to inflate the error term unnecessarily. We also want to use a separate error term for each analysis, rather than pooling these across Conditions.

The relevant analyses are presented in Table 14.9 for simple effects at one level of the other variable. Tests at the other levels would be carried out in the same way. Although this table has more simple effects than we care about, they are presented to illustrate the way in which tests were constructed. You would probably be foolish to consider all of the tests that result from this approach, because you would seriously inflate the familywise error rate. Decide what you want to look at before you run the analyses, and then stick to that decision. If you really want to look at a large number of simple effects, consider adopting one of the Bonferroni approaches discussed in Chapter 12.

Table 14.9 Analysis of simple effects

(a) Between-subjects effects (Condition, Sex, and Condition × Sex) at Pretest

Source	df	SS	MS	*F*
Condition	1	403.225	403.225	1.45
Sex	1	2544.025	2544.025	9.13*
Condition × Sex	1	119.025	119.025	0.43
Error	36	10027.100	278.530	
Total	39	13093.375		

(b) Within-subject effects (Sex, Time, Time × Sex) at BST

Source	df	SS	MS	*F*
Between subjects	19	7849.13		
Sex	1	2173.61	2173.61	6.89*
Error (between)	18	5675.52	315.30	
Within subjects	60	3646.26		
Time	3	338.94	112.98	1.88
T × S	3	54.54	18.18	0.30
Error (within)	54	3252.78	60.24	
Total	79	11495.39		

*$p < .05$

From the between-subjects analysis in Table14.9a we see that at Time 1 (Pretest) there was a significant difference between males and females (females show a lower frequency of use). But there were no Condition effects nor was there a Condition × Sex interaction. Males exceed females by just about the same amount in each Condition. The fact that there is no Condition effect is reassuring, because it would not be comforting to find that our two conditions differed before we had applied any treatment.

From the results in Table 14.9b we see that for the BST condition there is again a significant difference due to Sex, but there is no Time effect, nor a Time × Sex interaction. This is discouraging: It tells us that when we average across Sex there is no change in frequency of condom use as a result of our intervention. This runs counter to the conclusion that we might have drawn from the overall analysis where we saw a significant Condition by Time interaction, and speaks to the value of examining simple effects. The fact that an effect we seek is significant does not necessarily mean that it is significant in the direction we desire.

14.9. TWO WITHIN-SUBJECTS VARIABLES AND ONE BETWEEN-SUBJECTS VARIABLE

The design we just considered can be seen as a straightforward extension of the case of one between- and one within-subjects variable. All that we needed to add to the summary table was another main effect and the corresponding interactions. However, when we examine a design with two within-subjects main effects, the problem becomes slightly more complicated because

of the presence of additional error terms. To use a more generic notation, we will label the independent variables as *A*, *B,* and *C*.

Suppose that as a variation on the previous study we continued to use different subjects for the two levels of variable *A* (Gender), but we ran each subject under all combinations of variables *B* (Condition) and *C* (Trials). This design can be diagrammed as

	A_1			A_2		
	C_1	C_2	C_3	C_1	C_2	C_3
B_1	G_1	G_1	G_1	G_2	G_2	G_2
B_2	G_1	G_1	G_1	G_2	G_2	G_2
B_3	G_1	G_1	G_1	G_2	G_2	G_2

Before we consider an example, we will examine the expected mean squares for this design. These are presented in Table 14.10 for the case of the model in which all factors other than subjects are fixed. (subjects are treated as a random factor.) From the expected mean squares it is evident that we will have four error terms for this design. As before, the $MS_{\text{Ss w/in groups}}$ is used to test the between-subjects effect. When it comes to the within-subjects terms, however, *B* and the interaction of *B* with *A* are tested by *B* × *S*s within groups; *C* and its interaction with *A* are tested by *C* × *S*s within groups; and *BC* and its interaction with *A* are tested by *BC* × *S*s within groups. Why this is necessary is apparent from the expected mean squares

Table 14.10 Expected mean squares

Source	***df***	***E*(MS)**
Between subjects	$an - 1$	
A (groups)	$a - 1$	$\sigma_\varepsilon^2 + bc\sigma_\gamma^2 + nbc\theta_\alpha^2$
Ss w/in groups	$a(n - 1)$	$\sigma_\varepsilon^2 + bc\sigma_\gamma^2$

Within subjects	$na(bc-1)$	
B	$b-1$	$\sigma_\varepsilon^2+c\sigma_{\beta\pi}^2+nac\theta_\beta^2$
AB	$(a-1)(b-1)$	$\sigma_\varepsilon^2+c\sigma_{\beta\pi}^2+nc\theta_{\alpha\beta}^2$
$B\times$ Ss w/in groups	$a(b-1)(n-1)$	$\sigma_\varepsilon^2+c\sigma_{\beta\pi}^2$
C	$c-1$	$\sigma_\varepsilon^2+b\sigma_{\gamma\pi}^2+nab\theta_\gamma^2$
AC	$(a-1)(c-1)$	$\sigma_\varepsilon^2+b\sigma_{\gamma\pi}^2+nb\theta_{\alpha\gamma}^2$
$C\times$ Ss w/in groups	$a(c-1)(n-1)$	$\sigma_\varepsilon^2+b\sigma_{\gamma\pi}^2$
BC	$(b-1)(c-1)$	$\sigma_\varepsilon^2+n\sigma_{\beta\gamma\pi}^2+na\theta_{\beta\gamma}^2$
ABC	$(a-1)\,(b-1)(c-1)$	$\sigma_\varepsilon^2+n\sigma_{\beta\gamma\pi}^2+n\sigma_{\alpha\beta\gamma}^2$
$BC\times$ Ss w/in groups	$a(b-1)(c-1)(n-1)$	$\sigma_\varepsilon^2+n\sigma_{\beta\gamma\pi}^2$
Total	$N-1$	

An analysis of data on conditioned suppression

Assume that a tiny “click” on your clock radio always slightly precedes your loud and intrusive alarm going off. Over time that click (psychologists would call it a “CS”) could come to elicit the responses normally produced by the alarm (the “US”). Moreover, it is possible that simply presenting the click might lead to the suppression of an ongoing behavior, even if that click is not accompanied by the alarm. (If you were lying there reading, you might pause in your reading.) In a laboratory investigation of how the click affects (suppresses) ongoing behavior, Bouton and Swartzentruber (1985) investigated the degree to which a tone, which had previously been paired with shock, would suppress the rate of an ongoing bar-pressing response in rats. Suppression was measured by taking the ratio of the number of bar presses during a 1-minute test period following the tone to the total number of bar presses during both a baseline period and the test period. For all groups, behavior was assessed in two Phases—a Shock phase (shock accompanied the tone)

and a No-Shock phase (shock did not accompany the tone) repeated over a series of four Cycles of the experiment.

It may be easier to understand the design of the study if you first glance at the layout of Table 14.11. During Phase I, Group *A-B* was placed in Box *A*. After a 1-minute baseline interval, during which the animal bar-pressed for food, a tone was presented for 1 minute and was followed by a mild shock. The degree of suppression of the bar-pressing response when the tone was present (a normal fear response) was recorded. The animal was then placed in Box *B* for Phase II of the cycle, where, after 1 minute of baseline bar-pressing, only the tone stimulus was presented. Since the tone was previously paired with shock, it should suppress bar-pressing behavior to some extent. Over a series of *A-B* cycles, however, the subject should learn that shock is never administered in Phase II and that Box *B* is therefore a "safe" box. Thus, for later cycles there should be less suppression on the no-shock trials.

Group *L-A-B* was treated in the same way as Group *A-B* except that these animals previously had had experience with a situation in which a light, rather than a tone, had been paired with shock. Because of this previous experience, the authors expected the animals to perform slightly better (less suppression during Phase II) than did the other group, especially on the first cycle or two.

Group *A-A* was also treated in the same way as Group *A-B* except that both Phases were carried out in the same box—Box *A*. Because there were no differences in the test boxes to serve as cues (i.e., animals had no way to distinguish the no-shock from the shock phases), this group would be expected to show the most suppression during the No-shock phases.

Bouton and Swartzentruber predicted that overall there would be a main effect due to Phase (i.e., a difference between shock and no-shock Phases), a main effect due to Groups (*A-B* and *L-A-B* showing less suppression than *A-A*), and a main effect due to Cycles (animals tested in Box *B* would learn over time that it was a safe location). They also predicted that each of the interactions would be significant. (One reason I chose to use this example, even though it is difficult to describe concisely, is that it is one of those rare studies in which all effects are predicted to be significant and meaningful.)

The data and analysis of variance for this study are presented in Table 14.11. The analysis has not been elaborated in detail because it mainly involves steps that you already know how to do. The results are presented graphically in Figure 14.4 for convenience, and for the most part they are clear-cut and in the predicted direction. Keep in mind that for these data a lower score represents more suppression—that is, the animals are responding more slowly.

Table 14.11 Analysis of conditioned suppression (Lower scores represent greater suppression.)

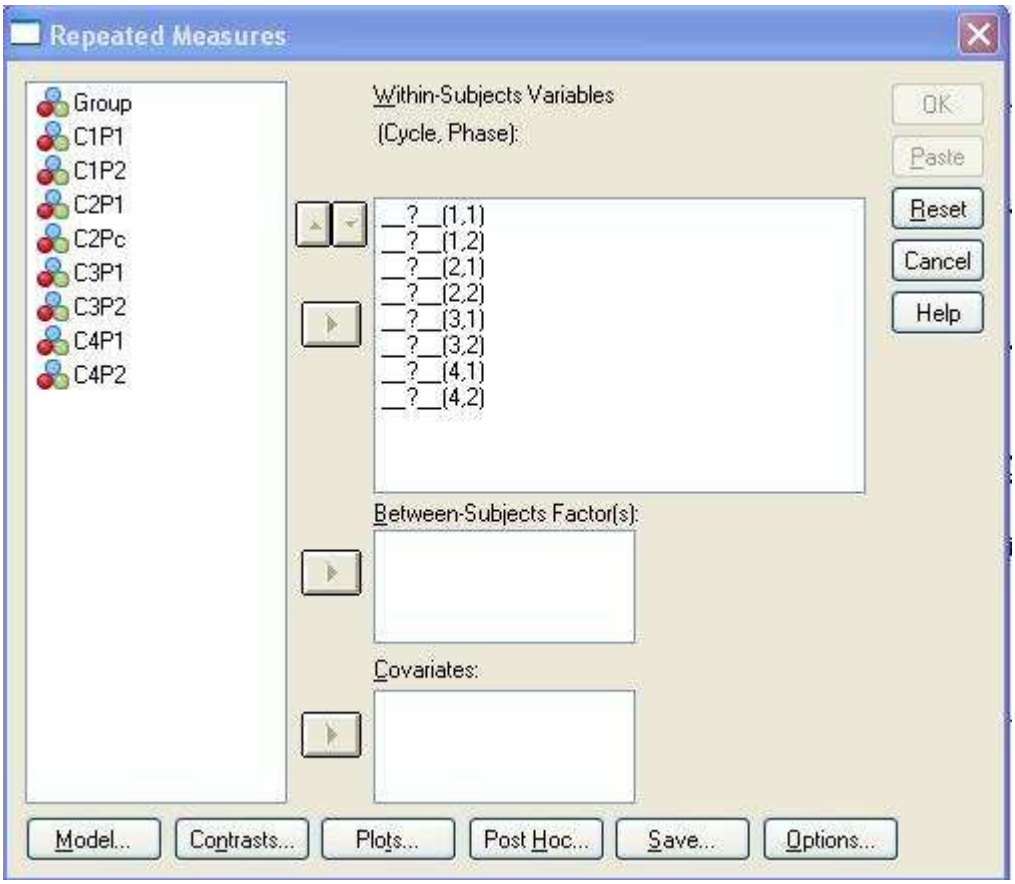
(a_1) Data

	Cycle								
	1		2		3		4		
	Phase		Phase		Phase		Phase		
Group	**I**	**II**	**I**	**II**	**I**	**II**	**I**	**II**	Subject Mean
A-B	1*	28	22	48	22	50	14	48	
									29.125
	21	21	16	40	15	39	11	56	
									27.375
	15	17	13	35	22	45	1	43	
									23.875
	30	34	55	54	37	57	57	68	
									49.000
	11	23	12	33	10	50	8	53	
									25.000
	16	11	18	34	11	40	5	40	
									21.875
	7	26	29	40	25	50	14	56	
									30.875
	0	22	23	45	18	38	15	50	
									26.375
$Mean_{AB}$	12.625	22.750	23.500	41.125	20.000	46.125	15.625	51.750	
									29.188
A-A	1	6	16	8	9	14	11	33	
									12.250
	37	59	28	36	34	32	26	37	
									36.125
	18	43	38	50	39	15	29	18	
									31.250
	1	2	9	8	6	5	5	15	
									6.375
	44	25	28	42	47	46	33	35	
									37.500
	15	14	22	32	16	23	32	26	
									22.500
	0	3	7	17	6	9	10	15	
									8.375
	26	15	31	32	28	22	16	15	
									23.125
$Mean_{AA}$	17.750	20.875	22.375	28.125	23.125	20.750	20.250	24.250	
									22.188
L-A-B	33	43	40	52	39	52	38	48	
									43.125
	4	35	9	42	4	46	23	51	
									26.750
	32	39	38	47	24	44	16	40	
									35.000
	17	34	21	41	27	50	13	40	
									30.375
	44	52	37	48	33	53	33	43	

									42.875
	12	16	9	39	9	59	13	45	
									25.250
	18	42	3	62	45	49	60	57	
									42.000
	13	29	14	44	9	50	15	48	
									27.750
$Mean_{LAB}$	21.625	36.250	21.375	46.875	23.750	50.375	26.375	46.500	
									34.141
Total	17.333	26.625	22.417	38.708	22.292	39.083	20.750	40.833	
									28.505

*Decimal points have been omitted in the table, but included in the calculations.

Rather than present literally three pages of tables and calculations, which few people would have the patience to work through, I have chosen to carry out the analysis using SPSS[6]. The data would be entered just as they appear in Table 14.11, with a column for Groups on the left. You would **select Analyze, General Linear Model, Repeated Measures** from the drop-down menus and specify that there were two repeated measures (Cycles with 4 levels and Phases with 2 levels). Then click on **Define** and specify the variables that are associated with each of the cells and the variable(s) that define the Between-Subject Factor(s). This dialogue box follows, where C1P1 – C4P2 would be moved to the Within-Subject Variables box and Group would be moved to the Between-Subjects Factor(s) box.



From the bottom row of that dialogue box you can specify what plots you would like to see, what contrasts you would like to run, and any descriptive statistics you want printed out. Then click on OK to run the analysis.

An abbreviated summary table appears below. I have omitted entries in the table related to Greenhouse and Geisser and related corrections to condense the table. Notice that SPSS presents separate tables for Within-Subject factors and Between-Subject factors, though I would prefer to have them combined into one table with appropriate indentations.

Table 14.12 SPSS output of the analysis of conditioned suppression data

Tests of Between-Subjects Effects

Measure: MEASURE_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	156009.005	1	156009.005	208.364	.000
Group	4616.760	2	2308.380	3.083	.067
Error	15723.359	21	748.731		

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Cycle	Sphericity Assumed	2726.974	3	908.991	12.027	.000
Cycle * Group	Sphericity Assumed	1047.073	6	174.512	2.309	.044
Error(Cycle)	Sphericity Assumed	4761.328	63	75.577		
Phase	Sphericity Assumed	11703.130	1	11703.130	129.855	.000
Phase * Group	Sphericity Assumed	4054.385	2	2027.193	22.493	.000
Error(Phase)	Sphericity Assumed	1892.609	21	90.124		
Cycle * Phase	Sphericity Assumed	741.516	3	247.172	4.035	.011
Cycle * Phase * Group	Sphericity Assumed	1273.781	6	212.297	3.466	.005
Error(Cycle*Phase)	Sphericity Assumed	3859.078	63	61.255		

Notice that there are multiple error terms in the table. The Group effect is tested by the Error term in the Between-Subjects table. Then Cycle and Cycle x Group are tested by Error(Cycle), Phase and Phase x Group are tested by Error(Phase), and Cycle x Phase and Cycle x Phase x Group are tested by Error(Cycle x Phase).

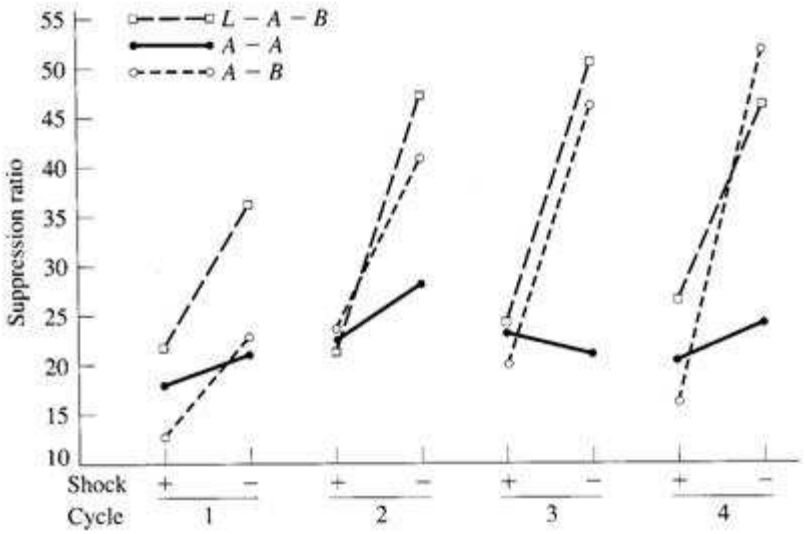


Figure 14.4 Conditioned suppression data

From the summary table in Table 14.12, it is clear that nearly all the predictions were supported. The only effect that was not significant was the main effect of Groups, but that effect is not crucial because it represents an average across the shock and the no-shock phases, and the experimenters had predicted little or no group differences in the shock phase. In this context, the Phase × Group interaction is of more interest, and it is clearly significant.

The presence of an interpretable three-way interaction offers the opportunity to give another example of the use of simple interaction effects. We would have predicted that all groups would show high levels of suppression of the shock trials on all Cycles, because anticipated shock is clearly disruptive. On no-shock trials, however, Groups *A-B* and *L-A-B* should show less suppression (higher scores) than Group *A-A*, and this latter difference should increase with Cycles. In other words, there should be a Groups × Cycles interaction for the no-shock trials, but no such interaction for the shock trials. The simple effects are shown in Table 14.13. (In these

tables Ihave left in the corrections based on Greenhouse-Geisser, Huhyn-Feldt, and Lower-Bound solutions to illustrate how they are presented by SPSS. Whether or not we choose to implement the corrections does not affect the conclusions. The calculation of the appropriate tests was carried out the same way it was earlier, by running a reduced analysis of variance using only the Phase 1 (or Phase 2) cells. Here again we are using separate error terms to test the Shock and No-shock effects, thus reducing problems with the sphericity assumption. (Again, just because the analyses also give simple effects due to Groups and Cycles is no reason to feel an obligation to interpret them. If they don't speak to issues raised by the experimental hypotheses, they should neither be reported nor interpreted unless you take steps to minimize the increase in the experimentwise error rate.)

Table 14.13 Simple interaction effects on conditioned suppression data

(a) Within-subject effects (Group × Cycle at Phase I)

Tests of Between-Subjects Effects

Measure: MEASURE_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	41126.760	1	41126.760	73.845	.000
Group	458.396	2	229.198	.412	.668
Error	11695.594	21	556.933		

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Cycle	Sphericity Assumed	403.615	3	134.538	1.740	.168
	Greenhouse-Geisser	403.615	2.391	168.788	1.740	.180
	Huynh-Feldt	403.615	2.977	135.598	1.740	.168
	Lower-bound	403.615	1.000	403.615	1.740	.201
Cycle * Group	Sphericity Assumed	415.604	6	69.267	.896	.504
	Greenhouse-Geisser	415.604	4.783	86.901	.896	.488
	Huynh-Feldt	415.604	5.953	69.813	.896	.503
	Lower-bound	415.604	2.000	207.802	.896	.423
Error(Cycle)	Sphericity Assumed	4871.031	63	77.318		
	Greenhouse-Geisser	4871.031	50.216	97.001		
	Huynh-Feldt	4871.031	62.508	77.927		
	Lower-bound	4871.031	21.000	231.954		

(b) Within-subject effects (Group × Cycle at Phase II)

Tests of Between-Subjects Effects

Measure: MEASURE_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	126585.375	1	126585.375	449.008	.000
Group	8212.750	2	4106.375	14.566	.000
Error	5920.375	21	281.923		

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Cycle	Sphericity Assumed	3064.875	3	1021.625	17.166	.000
	Greenhouse-Geisser	3064.875	2.275	1347.224	17.166	.000
	Huynh-Feldt	3064.875	2.809	1091.085	17.166	.000
	Lower-bound	3064.875	1.000	3064.875	17.166	.000
Cycle * Group	Sphericity Assumed	1905.250	6	317.542	5.336	.000
	Greenhouse-Geisser	1905.250	4.550	418.744	5.336	.001
	Huynh-Feldt	1905.250	5.618	339.131	5.336	.000
	Lower-bound	1905.250	2.000	952.625	5.336	.013
Error(Cycle)	Sphericity Assumed	3749.375	63	59.514		
	Greenhouse-Geisser	3749.375	47.774	78.481		
	Huynh-Feldt	3749.375	58.989	63.560		
	Lower-bound	3749.375	21.000	178.542		

From the simple interaction effects of Group × Cycle at each level of Phase, you can see that Bouton and Swartzentruber's predictions were upheld. There is no Cycle × Group interaction on Shock trials, but there is a clear interaction on No-shock trials.

14.10. INTRACLASS CORRELATION

One of the important issues in designing experiments in any field is the question of the reliability of the measurements. Most of you would probably expect that the *last* place to look for anything about reliability is in a discussion of the analysis of variance, but that is exactly where you will find it. (For additional material on the intraclass correlation, go to http://www.uvm.edu/~dhowell/StatPages/More_Stuff/icc/icc.html)

Suppose that we are interested in measuring the reliability with which judges rate the degree of prosocial behavior in young children. We might investigate this reliability by having two or more judges each rate a behavior sample of a number of children, assigning a number from 1 to 10 to reflect the amount of prosocial behavior in each behavior sample. I will demonstrate the procedure with some extreme data that were created to make a point. Look at the data in Table 14.18.

Table 14.18

	(a) Judge			(b) Judge			(c) Judge		
Child	I	II	III	I	II	III	I	II	III
1	1	1	2	1	0	3	1	3	7
2	3	3	3	3	2	5	3	1	5
3	5	5	5	5	4	7	5	7	4
4	5	6	6	5	4	7	5	5	5
5	7	7	7	7	6	8	7	6	7

In Table 14.18a the judges are in almost perfect agreement. They all see wide differences between children, they all agree on which children show high levels of prosocial behavior and which show low levels, *and* they are nearly in agreement on how high or low those levels are. In this case nearly all of the variability in the data involves differences among children—there is almost no variability among judges and almost no random error.

In Table 14.18b we see much the same pattern, but with a difference. The judges do see overall differences among the children, and they do agree on which children show the highest (and lowest) levels of the behavior. But the judges disagree in terms of the amount of prosocial behavior they see. Judge II sees slightly less behavior than Judge I (his mean is 1 point lower), and Judge III sees relatively more behavior than do the others. In other words, while the judges agree on *ordering* children, they disagree on *level*. Here the data involve both variability among children and variability among judges. However, the random error component is still very small. This is often the most realistic model of how people rate behavior because each of us has a different understanding of how much behavior is required to earn a rating of "7," for example. Our assessment of the reliability of a rating system must normally take variability among judges into account.

Finally, Table 14.18c shows a pattern where not only do the judges disagree in level, they also disagree in ordering children. A large percentage of the variability in these data is error variance.

So what do we do when we want to talk about reliability? One way to measure reliability when judges use only a few levels or categories is to calculate the percentage of times that two judges agree on their rating, but this measure is biased because of high levels of chance agreement whenever one or two categories predominate. (But see the discussion earlier of Cohen's kappa.) Another common approach is to correlate the ratings of two judges, and perhaps average pairwise correlations if you have multiple judges. But this approach will not take differences between judges into account. (If one judge always rates five points higher than another judge the correlation will be 1.00, but the judges are saying different things about the subjects.) A third way is to calculate what is called the **intraclass correlation**, taking differences due to judges into account. That is what we will do here.

You can calculate an intraclass correlation coefficient in a number of different ways, depending on whether you treat judges as a fixed or random variable and whether judges evaluate the same or different subjects. The classic reference for intraclass correlation is Shrout and Fleiss (1979), who discuss several alternative approaches. I am going to discuss only the most common approach here, one in which we consider our judges to be a random sample of all judges we could have used and in which each judge rates the same set of subjects once. (In what follows I am assuming that judges are rating "subjects," but they could be rating pictures, cars, or the livability of cities. Take the word "subject" as a generic term for whatever is being rated.)

We will start by assuming that the data in Table 14.18 can be represented by the following model:

$$X_{ij} = \mu + \alpha_i + \pi_j + \alpha\pi_{ij} + e_{ij}$$

In this model α_i stands for the effect of the *i*th judge, π_j stands for the effect of the *j*th subject (person), $\alpha\pi_{ij}$ is the interaction between the *i*th judge and the *j*th subject (the degree to which the judge changes his or her rating system when confronted with that subject), and e_{ij} stands for the error associated with that specific rating. Because each judge rates each subject only once, it is not possible in this model to estimate $\alpha\pi_{ij}$ and e_{ij} separately, but it is necessary to keep them separate in the model.

If you look back to the previous chapter you will see that when we calculated a magnitude-of-effect measure (which was essentially an r^2-family measure), we took the variance estimate for the effect in question (in this case differences among subjects) relative to the sum of the estimates of the several sources of variance. That is precisely what we are going to do here. We will let

$$\text{Intraclass correlation} = \sigma_s^2 / \left(\sigma_\alpha^2 + \sigma_s^2 + \sigma_{\alpha s}^2 + \sigma_e^2\right)$$

If most of the variability in the data is due to differences between subjects, with only a small amount due to differences between judges, the interaction of judges and subjects, and error, then this ratio will be close to 1.00. If judges differ from one another in how high or low they rate people in general, or if there is a judge by subject interaction (different judges rate different

people differently), or if there is a lot of error in the ratings, the denominator will be substantially larger than the numerator and the ratio will be much less than 1.00.

To compute the intraclass correlation we are first going to run a Subjects × Judges analysis of variance with Judges as a repeated measure. Because each judge rates each subject only once, there will not be an independent estimate of error, and we will have to use the Judge × Subject interaction as the error term. From the summary table that results, we will compute our estimate of the intraclass correlation as

$$\text{Intraclass correlation} = \frac{MS_{\text{Subjects}} - MS_{J\times S}}{MS_{\text{Subjects}} + (j-1)MS_{J\times S} + j\left(MS_{\text{Judge}} - MS_{J\times S}\right)/n}$$

where *j* represents the number of judges and *n* represents the number of subjects.

To illustrate this, I have run the analysis of variance on the data in Table14.18b, which is the data set where I have deliberately built in some differences due to subjects and judges. The summary table for this analysis follows.

Source	*df*	SS	MS	*F*
Between subjects	4	57.067	14.267	
Within subjects	10	20.666	2.067	
Judge	2	20.133	10.067	150.25
Judge × Subjects	8	0.533	0.067	
Total	14	77.733		

We can now calculate the intraclass correlation as

$$\text{Intraclass correlation} = \frac{14.267 - 0.067}{14.267 + (3-1)0.067 + 3(10.067 - 0.067)/5}$$

$$= \frac{14.200}{14.267 + 0.134 + 6} = \frac{14.2}{20.401} = .70$$

Thus our measure of reliability is .70, which is probably not as good as we would like to see it. But we can tell from the calculation that the main thing that contributed to low reliability was not error, but differences among judges. This would suggest that we need to have our judges work together to decide on a consistent scale where a "7" means the same thing to each judge.

14.11. OTHER CONSIDERATIONS

Sequence effects

Repeated-measures designs are notoriously susceptible to **sequence effects** and **carryover** (practice) **effects**. Whenever the possibility exists that exposure to one treatment will influence the effect of another treatment, the experimenter should consider very seriously before deciding to use a repeated-measures design. In certain studies, carryover effects are desirable. In learning studies, for example, the basic data represent what is carried over from one trial to another. In most situations, however, carryover effects (and especially differential carryover effects) are considered a nuisance—something to be avoided.

The statistical theory of repeated-measures designs assumes that the order of administration is randomized separately for each subject, unless, of course, the repeated measure is something like trials, where it is impossible to have trial 2 before trial 1. In some situations, however, it makes more sense to assign testing sequences by means of a **Latin square** or some other device. Although this violates the assumption of randomization, in some situations the gains outweigh the losses. What is important, however, is that random assignment, Latin squares, and so on do

not in themselves eliminate sequence effects. Ignoring analyses in which the data are *analyzed* by means of a Latin square or a related statistical procedure, any system of assignment simply distributes sequence and carryover effects across the cells of the design, with luck lumping them into the error term(s). The phrase "with luck" implies that if this does not happen, the carryover effects will be confounded with treatment effects and the results will be very difficult, if not impossible, to interpret. For those students particularly interested in examining sequence effects, Winer (1971), Kirk (1968), and Cochran and Cox (1957) present excellent discussions of Latin square and related designs.

Unequal group sizes

One of the pleasant features of repeated-measures designs is that when a subject fails to arrive for an experiment, it often means that that subject is missing from every cell in which he was to serve. This has the effect of keeping the cell sizes proportional, even if unequal. If you are so unlucky as to have a subject for whom you have partial data, the common procedure is to eliminate that subject from the analysis. If, however, only one or two scores are missing, it is possible to replace them with estimates, and in many cases this is a satisfactory approach. For a discussion of this topic, see Federer (1955, pp. 125–126, 133ff), and especially Little and Rubin (1987), and Howell (2008) and the discussion in Section 14.12.

Matched samples and related problems

In discussing repeated-measures designs, we have spoken in terms of repeated measurements on the same subject. Although this represents the most common instance of the use of these designs, it is not the only one. The specific fact that a subject is tested several times really has nothing to do with the matter. Technically, what distinguishes repeated-measures designs (or, more generally, **randomized blocks designs**, of which repeated-measures designs are a special case) from the common factorial designs with equal *n*s is the fact that for repeated-measures designs, the off-diagonal elements of $\sum$ do not have an expectation of zero—that is, the treatments are correlated. Repeated use of the same subject leads to such correlations, but so does use of **matched samples** of subjects. Thus, for example, if we formed 10 sets of three subjects each, with the subjects matched on driving experience, and then set up an experiment in which the first subject under each treatment came from the same matched triad, we would have correlations among treatments and would thus have a repeated-measures design. Any other data-collection procedure leading to nonzero correlations (or covariances) could also be treated as a repeated-measures design.

14.12. MIXED MODELS FOR REPEATED-MEASURES DESIGNS

Earlier in the chapter I said that the standard repeated-measures analysis of variance requires an assumption about the variance–covariance matrix known as *sphericity*, a specific form of which is known as *compound symmetry*. When we discussed $\hat{\varepsilon}$ and $\tilde{\varepsilon}$, we were concerned with

correction factors that we could apply to the degrees of freedom to circumvent some of the problems associated with a failure of the sphericity assumption.

There is a considerable literature on repeated-measures analyses and their robustness in the face of violations of the underlying assumptions. Although there is not universal agreement that the adjustments proposed by Greenhouse and Geisser and by Huynh and Feldt are successful, the adjustments work reasonably well as long as we focus on overall main or interaction effects, or as long as we use only data that relate to specific simple effects (rather than using overall error terms). Where we encounter serious trouble is when we try to run individual contrasts or simple effects analyses using pooled error terms. Boik (1981) has shown that in these cases the repeated-measures analysis is remarkably sensitive to violations of the sphericity assumption unless we adopt separate error terms for each contrast, as I did for the simple effects tests in Table 14.13. However there is another way of dealing with assumptions about the covariance matrix, and that is to not make such assumptions. But to do that we need to take a different approach to the analysis itself.

Standard repeated measures analysis of variance has two problems that we have lived with for many years and will probably continue to live with. It assumes both compound symmetry (or sphericity) and complete data. If a participant does not appear for a follow-up session, even if he appears for all of the others, he must be eliminated from the analysis. There is an alternative approach to the analysis of repeated measures designs that does not hinge on either sphericity assumptions or complete data. This analysis is often referred to as **mixed models**, **multilevel modeling**, or **hierarchical modeling**. There is a bit of confusion here because we have already

used the phrase “mixed models” to refer to any experimental design that involves both fixed and random factors. That is a perfectly legitimate usage. But when we are speaking of a method of analysis, such as we are here, the phrase “mixed models” refers more to a particular type of solution, involving both fixed and random factors, using a different approach to the arithmetic. More specifically, when someone claims to have done their analysis using mixed models, they are referring to a solution that employs **maximum likelihood** or, more likely, **restricted maximum likelihood (REML)** in place of the least squares approaches that we have focused on up to now and will focus on again in the next two chapters[7].

This section covers a small part of the broader topic of hierarchical or multilevel models. For these models the repeated measure (e.g. Time or Trials) is a fixed factor while Subjects is a random factor. The between-subjects factor is also usually a fixed factor. By approaching the problem using restricted maximum likelihood (REML) as the method of parameter estimation, the solution can take cognizance from the very beginning of the analysis that one or more factors are fixed and one or more factors are random. Least squares solutions of standard analysis of variance treats all factors as fixed until it comes to determining error terms for *F* statistics.

No one would seriously attempt to do employ a mixed model analysis by hand. You must use computer software to perform the analysis. However there are many software programs available, some of them even free. The ones that you will have most access to are probably **SPSS Mixed** and **SAS Proc Mixed**. I will use SPSS for our example, though **SAS proc mixed** is probably more flexible. A more complete discussion of the analysis of alternative designs can be found at http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Mixed Models for

Repeated Measures.pdf For an example I have chosen a design with one between subject variable and one within subject variable. The example has missing data because that will illustrate an analysis that you can not do with standard analysis of variance.

The Data

I created data to have a number of characteristics. There are two groups – a Control group and a Treatment group, measured at 4 times. These times are labeled as 0 (pretest), 1 (one month posttest), 3 (three months follow-up), and 6 (six months follow-up). I had a study of treatment of depression in mind, so I created the treatment group to show a sharp drop in depression at post-test and then sustain that drop (with slight regression) at 3 and 6 months. The Control group declines slowly over the 4 intervals but does not reach the low level of the Treatment group.

The data are shown in Table 14.19. A period is used to indicate missing values.

Table 14.19 Data for mixed model analysis.

Group	Subj	Time0	Time1	Time3	Time6
1	1	296	175	187	242
1	2	376	329	236	126
1	3	309	238	150	173
1	4	222	60	82	135
1	5	150	.	250	266
1	6	316	291	238	194
1	7	321	364	270	358
1	8	447	402	.	266
1	9	220	70	95	137
1	10	375	335	334	129
1	11	310	300	253	.
1	12	310	245	200	170

Group	Subj	Time0	Time1	Time3	Time6
2	13	282	186	225	134
2	14	317	31	85	120
2	15	362	104	.	.
2	16	338	132	91	77
2	17	263	94	141	142
2	18	138	38	16	95
2	19	329	.	.	6
2	20	292	139	104	.
2	21	275	94	135	137
2	22	150	48	20	85
2	23	319	68	67	.
2	24	300	138	114	174

One difference between data files for mixed models and others is that we use what is often called a "long form." Instead of putting each subject's data on one line, we have a separate line for every value of the dependent variance. Thus our data file will be structured like the one in Table 14.20

Table 14.20 Data restructured into a long form.

Subj	Time	Group	dv
1	0	1	296
1	1	1	175
1	3	1	187
1	6	1	242
… …	…	…	
24	3	2	114
24	6	2	174

Instead of showing you how to use the graphical interface in SPSS, which would take quite a bit of space, I am simply giving you the syntax for the commands[8]. After you have entered your data, open a new Syntax window, paste in the following commands, and select Run from the toolbar. I have left out a number of commands that do fine tuning, but what I have will run your analysis nicely.

```
MIXED
```

```
dv  BY Group Time
/FIXED = Group Time Group*Time  | SSTYPE(3)
/METHOD = REML
/PRINT = DESCRIPTIVES SOLUTION
/REPEATED = Time | SUBJECT(Subj) COVTYPE(CS)
/EMMEANS = TABLES(Group)
/EMMEANS = TABLES(Time)
/EMMEANS = TABLES(Group*Time)  .
```

I am only presenting the most important parts of the printout, but you can see the rest by running the analysis yourself. (The data are available on the book's website as WickMiss.dat.)

Information Criteria[a]

-2 Restricted Log Likelihood	905.398
Akaike's Information Criterion (AIC)	909.398
Hurvich and Tsai's Criterion (AICC)	909.555
Bozdogan's Criterion (CAIC)	916.136
Schwarz's Bayesian Criterion (BIC)	914.136

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: dv.

Fixed Effects

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	22.327	269.632	.000
Group	1	22.327	16.524	.001
Time	3	58.646	32.453	.000
Group * Time	3	58.646	6.089	.001

a. Dependent Variable: dv.

Covariance Parameters

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error
Repeated Measures	CS diagonal offset	2954.544	551.1034
	CS covariance	2558.656	1026.581

a. Dependent Variable: dv.

I will not discuss the section labeled “Information criteria” here, but will come back to it when we compare the fit of different models. The fixed effects part of the table looks just like one that you would see in most analyses of variance except that it does not include sums of squares and mean squares. That is because of the way that maximum likelihood solutions go about solving the problem. In some software it is possible to force them into the printout. Notice the test on the Intercept. That is simply a test that the grand mean is 0, and is of no interest to us. The other three effects are all significant. We don’t really care very much about the two main effects. The groups started off equal on pre-test, and those null differences would influence any overall main effect of groups. Similarly, we don’t care a great deal about the Time effect because we expect different behavior from the two groups. What we do care about, however, is the interaction. This tells us that the two groups perform differently over Time, which is what we hoped to see. You can see this effect in Figure 14.5.

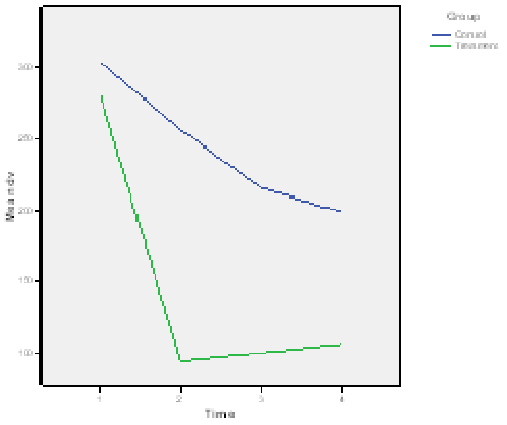


Figure 14.5 Means across trials for the two conditions.

There are two additional results in the printout that need to be considered. The section headed "Covariance Parameters" is the random part of the model. The term labeled "CS diagonal offset" represents the residual variance and, with balanced designs, would be the error term for the within-subject tests. The term labeled "CS covariance" is the variance of the intercepts, meaning that if you plot the dependent variable against time for each subject, the differences in intercepts of those lines would represent differences due to subjects (some lines are higher than others) and it is this variance that we have here. For most of us that variance is not particularly important, but there are studies in which it is.

As I said earlier, mixed model analyses do not require an assumption of compound symmetry. In fact, that assumption is often incorrect. In Table 14.21 you can see the pattern of correlations among trials. These are averaged over the separate groups, but give you a clear picture that the structure is not one of compound symmetry.

Table 14.21 Correlations among trials

Estimated R Correlation Matrix for Subject 1

Row	Col1	Col2	Col3	Col4
1	1.0000	0.5121	0.4163	-0.08840
2	0.5121	1.0000	0.8510	0.3628
3	0.4163	0.8510	1.0000	0.3827
4	-0.08840	0.3628	0.3827	1.0000

There are a number of things that we could do to alter the model that we just ran, which requested a solution based on compound symmetry. We could tell SPSS to solve the problem without assuming anything about the correlations or covariances. (That is essentially what the MANOVA approach to repeated measures does.) The problem with this approach is that the solution has to derive estimates of those correlations and that will take away degrees of freedom, perhaps needlessly. There is no point in declaring that you are totally ignorant when you are really only partially ignorant. Another approach would be to assume a specific (but different) form of the covariance matrix. For example, we could use what is called an autoregressive solution. Such a solution assumes that correlations between observations decrease as the times move further apart in time. It further assume that each correlation depends only on the preceding correlation plus some (perhaps much) error. If the correlation between adjacent trials is, for example 0.5121 (as it is in the study we are discussing), then times that are two steps apart are assumed to correlate $.5121^2$ and times three steps apart are assumed to correlate $.5121^3$. This leads to a matrix of correlations that decrease regularly the more removed the observations are from each other. That sounds like a logical expectation for what we would find when we measure depression over time. For now we are going to consider the autoregressive covariance structure.

Having decided on a correlational (or covariance) structure we simply need to tell SPSS to use that structure and solve the problem as before. The only change we will make is to the **repeated** command, where we will replace covtype(cs) with covtype(AR1).

```
MIXED
  dv  BY Group Time
  /FIXED = Group Time Group*Time  | SSTYPE(3)
  /METHOD = REML
  /PRINT = DESCRIPTIVES SOLUTION
  /REPEATED = Time | SUBJECT(Subj) COVTYPE(AR1)
  /EMMEANS = TABLES(Group)
  /EMMEANS = TABLES(Time)
  /EMMEANS = TABLES(Group*Time)  .
```

Information Criteria[a]

-2 Restricted Log Likelihood	895.066
Akaike's Information Criterion (AIC)	899.066
Hurvich and Tsai's Criterion (AICC)	899.224
Bozdogan's Criterion (CAIC)	905.805
Schwarz's Bayesian Criterion (BIC)	903.805

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: dv.

Fixed Effects

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	26.462	270.516	.000
Group	1	26.462	17.324	.000
Time	3	57.499	30.821	.000
Group * Time	3	57.499	7.721	.000

a. Dependent Variable: dv.

Covariance Parameters

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error
Repeated Measures	AR1 diagonal	5349.876	1060.035
	AR1 rho	.618198	.084130

a. Dependent Variable: dv.

Here we see that all effects are still significant, which is encouraging. But which of these two models (one assuming a compound symmetry structure to the covariance matrix and the other assuming a first order autoregressive structure) is the better choice. We are going to come to the same conclusion with either model in this case, but that is often not true, and we still want to know which model is better. One way of doing that is to compare the sections labeled "Information Criteria" for each analysis. These are reproduced below for the two models.

Compound Symmetry

Information Criteria[a]

-2 Restricted Log Likelihood	905.398
Akaike's Information Criterion (AIC)	909.398
Hurvich and Tsai's Criterion (AICC)	909.555
Bozdogan's Criterion (CAIC)	916.136
Schwarz's Bayesian Criterion (BIC)	914.136

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: dv.

Autoregressive (1)

Information Criteria[a]

-2 Restricted Log Likelihood	895.066
Akaike's Information Criterion (AIC)	899.066
Hurvich and Tsai's Criterion (AICC)	899.224
Bozdogan's Criterion (CAIC)	905.805
Schwarz's Bayesian Criterion (BIC)	903.805

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: dv.

A good way to compare models is to compare either the Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC). In general a model with a smaller value is better. For our examples the two AIC criteria are 909.398 and 899.066. It would appear that the Autoregressive (1) model is to be preferred, which is in line with what our eyes told us about the covariance structures. (If we had rerun the analysis using an unstructured covariance matrix (COVTYPE(UN)), AIC would be 903.691 and BIC would be 927.385, so we would still choose the autoregressive model.)

Mixed models have a great deal to offer in terms of fitting data to models and allow us to compare underlying models to best interpret our data. They also can be very valuable in the absence of missing data. However they are more difficult to work with and the software, while certainly improving, is far from intuitive in some cases. However I think that this is the direction that more and more analyses will take over the next decade and it is important to understand them.

Papers by Overall, Tonidandel, and others that illustrate the problems with mixed models. The major problem is the fact that it is very difficult to know how to correctly specify your model, and different specifications can lead to different results and sometimes rather low power. An excellent paper in this regard is by Overall and Shivakumar (1997) and another by Overall and Tonidandel (2007). I recommend that you look at those papers when considering the use of mixed models, although those authors used SAS **Proc Mixed** for their analyses and it is not entirely clear how those models relate to models you would have using SPSS. What seems to be

critically important is the case where missing data depend on the participant's initial response at baseline and attempts to use this measure as a covariate..

Key Terms

Partition (Introduction)

Partialling out (Introduction)

Repeated-measures designs (Introduction)

$SS_{between\ subj}$ **(**SS_s**)** (Introduction)

$SS_{within\ subj}$ (Introduction)

Main diagonal (14.3)

Off-diagonal elements (14.3)

Compound symmetry (14.3)

Covariance matrix ($\sum$**)** (14.3)

Sphericity (14.3)

Multivariate analysis of variance (MANOVA) (14.3)

Multivariate procedure (14.3)

$Error_{between}$ (14.7)

$Error_{within}$ (14.7)

Intraclass correlation (14.11)

Sequence effects (14.12)

Carryover effects (14.12)

Latin square (14.12)

Randomized blocks designs (14.12)

Matched samples (14.12)

Univariate (14.14)

EXERCISES

14.1) It is at least part of the folklore that repeated experience with any standardized test leads to better scores, even without any intervening study. We obtain eight subjects and give them a standardized admissions exam every Saturday morning for 3 weeks. The data follow:

S	First	Second	Third
1	550	570	580
2	440	440	470
3	610	630	610
4	650	670	670
5	400	460	450
6	700	680	710
7	490	510	510
8	580	550	590

a) Write the statistical model for these data.

b) Run the analysis of variance.

c) What, if anything, would you conclude about practice effects on the GRE?

14.2) Using the data from Exercise 14.1,

a) delete the data for the third session and run a (matched-sample) *t* test between Sessions 1 and 2.

b) Now run a repeated-measures analysis of variance on the two columns you used in part (a) and compare this *F* with the preceding *t*.

14.3) To demonstrate the practical uses of basic learning principles, a psychologist with an interest in behavior modification collected data on a study designed to teach self-care skills to severely developmentally handicapped children. An experimental group received reinforcement for activities related to self-care. A second group received an equivalent amount of attention, but no reinforcement. The children were scored (blind) by a rater on a 10-point scale of self-sufficiency. The ratings were done in a baseline session and at the end of training. The data follow:

Reinforcement		**No Reinforcement**	
Baseline	**Training**	**Baseline**	**Training**
8	9	3	5
5	7	5	5
3	2	8	10
5	7	2	5
2	9	5	3
6	7	6	10
5	8	6	9
6	5	4	5
4	7	3	7
4	9	5	5

Run the appropriate analysis and state your conclusions.

14.4) An experimenter with only a modicum of statistical training took the data in Exercise 14.3 and ran an independent-groups *t* test instead, using the difference scores (training minus baseline) as the raw data.

a) Run that analysis.

b) Square the value of *t* and compare it to the *F*s you obtained in Exercise 14.3.

c) Explain why t^2 is not equal to *F* for Groups.

14.5) To understand just what happened in the experiment involving the training of severely developmentally handicapped children (Exercise 14.3), our original experimenter evaluated a third group at the same times as he did the first two groups, but otherwise provided

no special treatment. In other words, these children did not receive reinforcement, or even the extra attention that the control group did. Their data follow:

Baseline:	3	5	8	5	5	6	6	6	3	4
Training:	4	5	6	6	4	7	7	3	2	2

a) Add these data to those in Exercise 14.3 and rerun the analysis.

b) Plot the results.

c) What can you conclude from the results you obtained in parts (a) and (b)?

d) Within the context of this three group experiment, run the contrast of the two conditions that you have imported from Exercise 14.3.

e) Compute the effect size for the contrast in part d).

14.6) For 2 years I carried on a running argument with my daughter concerning hand calculators. She wanted one. I maintained that children who use calculators never learn to do arithmetic correctly, whereas she maintained that they do. To settle the argument, we selected five of her classmates who had calculators and five who did not, and made a totally unwarranted assumption that the presence or absence of calculators was all that distinguished these children. We then gave each child three 10-point tests (addition, subtracton, and multiplication), which they were required to do in a very short time in their heads. The scores are as follows:

	Addition	**Subtraction**	**Multiplication**
Calculator owners	8	5	3
	7	5	2
	9	7	3
	6	3	1
	8	5	1
Non-calculator owners	10	7	6
	7	6	5
	6	5	5
	9	7	8
	9	6	9

a) Run the analysis of variance.

b) Do the data suggest that I should have given in and bought my daughter a calculator? (I did anyway. She is now in her late 30s and is a fully certified actuary—so what do I know?)

14.7) For the data in Exercise 14.6,

a) calculate the variance–covariance matrices.

b) calculate $\hat{\Sigma}$ using your answers to part (a).

14.8) From the results in Exercise 14.7, do we appear to have reason to believe that we have met the assumptions required for the analysis of repeated measures?

14.9) For the data in Exercise 14.6,

a) calculate all possible simple effects after first plotting the results.

b) test the simple effects, calculating test terms and adjusted degrees of freedom where necessary.

14.10) In a study of the way children and adults summarize stories, we selected 10 fifth graders and 10 adults. These were further subdivided into equal groups of good and poor readers (on the hypothesis that good and poor readers may store or retrieve story information differently). All subjects read 10 short stories and were asked to summarize the story in their own words immediately after reading it. All summaries were content analyzed, and the numbers of statements related to Settings, Goals, and inferred Dispositions were recorded. The data are collapsed across the 10 stories:

Age	**Adults**			**Children**		
Items	**Setting**	**Goal**	**Disp.**	**Setting**	**Goal**	**Disp.**
Good readers	8	7	6	5	5	2
	5	6	4	7	8	4
	5	5	5	7	7	4
	7	8	6	6	4	3
	6	4	4	4	4	2
Poor readers	7	6	3	2	2	2
	5	3	1	2	0	1
	6	6	2	5	4	1

4	4	1	4	4	2
5	5	3	2	2	0

Run the appropriate analysis.

14.11) Refer to Exercise 14.10.

a) Calculate the simple effect of reading ability for children.

b) Calculate the simple effect of items for adult good readers.

14.12) Calculate the within-groups covariance matrices for the data in Exercise 14.10.

14.13) Suppose we had instructed our subjects to limit their summaries to 10 words. What effect might that have on the data in Exercise 14.10?

14.14) In an investigation of cigarette smoking, an experimenter decided to compare three different procedures for quitting smoking (tapering off, immediate stopping, and aversion therapy). She took five subjects in each group and asked them to rate (on a 10-point scale) their desire to smoke "right now" in two different environments (home versus work) both before and after quitting. Thus, we have one between-subjects variable (Treatment group) and two within-subjects variables (Environment and Pre/Post).

	Pre		**Post**	
	Home	**Work**	**Home**	**Work**
Taper	7	6	6	4
	5	4	5	2
	8	7	7	4
	8	8	6	5
	6	5	5	3
Immediate	8	7	7	6
	5	5	5	4
	7	6	6	5
	8	7	6	5
	7	6	5	4
Aversion	9	8	5	4

4	4	3	2
7	7	5	3
7	5	5	0
8	7	6	3

a) Run the appropriate analysis of variance.

b) Interpret the results.

14.15) Plot the results you obtained in Exercise 14.14.

14.16) Run simple effects on the data in Exercise 14.14 to clarify the results.

14.17) The abbreviated printout in Exhibit 14.3 represents the analysis of the data in Exercise 14.5.

a) Compare this printout with the results you obtained in Exercise 14.5.

b) What does a significant *F* for "MEAN" tell us?

c) Relate $MS_{wincell}$ to the table of cell standard deviations.

Exhibit 14.3

BMDP2V - ANALYSIS OF VARIANCE AND COVARIANCES

WITH REPEATED MEASURES.

```
PROGRAM CONTROL INFORMATION
/PROBLEM      TITLE IS 'BMDP2V ANALYSIS OF EXERCISE 14.5'.
/INPUT        VARIABLES ARE 3.
              FORMAT IS '(3F2.0)'.
              CASES ARE 30.
/VARIABLE     NAMES ARE GROUP,  PRE,  POST.
/DESIGN       DEPENDENT ARE 2, 3.
              LEVELS ARE 2.
              NAME IS TIME.
              GROUP = 1.
/END
```

CELL MEANS FOR 1-ST DEPENDENT VARIABLE

					MARGINAL
GROUP	=	* 1.0000	* 2.0000	*3.0000	
	TIME				
PRE	1	4.80000	4.70000	5.10000	4.86667

POST	2	7.00000	6.40000	4.60000	6.00000
MARGINAL		5.90000	5.55000	4.85000	5.43333
COUNT		10	10	10	30

STANDARD DEVIATIONS FOR 1-ST DEPENDENT VARIABLE

GROUP =	* 1.0000	* 2.000	* 3.0000
TIME			
PRE 1	1.68655	1.76698	1.52388
POST 2	2.16025	2.45855	1.89737

	SOURCE	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F	TAIL PROBABILITY
	MEAN	1771.26667	1	1771.26667	322.48	0.0000
	GROUP	11.43333	2	5.71667	1.04	0.3669
1	ERROR	148.30000	27	5.49259		
	TIME	19.26667	1	19.26667	9.44	0.0048
	TG	20.63333	2	10.31667	5.06	0.0137
2	ERROR	55.10000	27	2.04074		

14.18) The SPSS printout in Exhibit 14.4 was obtained by treating the data in Exercise 14.10 as though all variables were between-subjects variables (i.e., as though the data represented a standard three-way factorial). Show that the error terms for the correct analysis represent a partition of the error term for the factorial analysis.

Exhibit 14.4

Tests of Between-Subjects Effects

Dependent Variable: DV

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	170.800[a]	11	15.527	9.001	.000
Intercept	1058.400	1	1058.400	613.565	.000
AGE	68.267	1	68.267	39.575	.000
READTYPE	29.400	1	29.400	17.043	.000
PART	60.400	2	30.200	17.507	.000
AGE * READTYPE	3.267	1	3.267	1.894	.175
AGE * PART	.933	2	.467	.271	.764
READTYPE * PART	.000	2	.000	.000	1.000
AGE * READTYPE * PART	8.533	2	4.267	2.473	.095
Error	82.800	48	1.725		
Total	1312.000	60			
Corrected Total	253.600	59			

a. R Squared = .674 (Adjusted R Squared = .599)

14.19) Outline the summary table for an $A \times B \times C \times D$ design with repeated measures on *A* and *B* and independent measures on *C* and *D*.

14.20) Foa, Rothbaum, Riggs, and Murdock (1991) ran a study comparing different treatments for posttraumatic stress disorder (PTSD). They used three groups (plus a waiting list control) One group received Stress Inoculation Therapy (SIT), another received a Prolonged Exposure (PE) treatment, and a third received standard Supportive Counseling (SC). All clients were measured at Pretreatment, Posttreatment, and a 3.5 month Follow-up. The data below closely approximate the data that they collected, and the dependent variable is a measure of PTSD.

SIT			PE			SC		
Pre	**Post**	**Followup**	**Pre**	**Post**	**Followup**	**Pre**	**Post**	**Followup**
19	6	1	20	5	0	12	14	18
28	14	16	21	18	21	27	18	9
18	6	8	36	26	17	24	19	13

23	6	11	25	11	9	32	21	11
21	6	13	26	2	7	26	20	18
24	10	8	30	31	10	18	20	26
26	10	7	19	6	11	38	35	34
15	6	13	19	7	5	26	22	22
18	8	6	22	4	4	23	10	8
34	13	8	22	17	20	22	19	19
20	10	16	24	19	1	34	27	23
34	10	1	28	22	16	22	15	12
29	16	23	29	23	20	27	18	13
33	19	39	27	15	20	23	21	19
22	7	16	27	7	3	26	18	13

a) Run a repeated measures analysis of variance on these data.

b) Draw the appropriate conclusions.

14.21) Using the data from Exercise 14.20 use SPSS to run a mixed models analysis of variance, specifying an appropriate form for the covariance matrix, and compare the results with those you obtained in Exercise 14.20.

14.22) The following data come from Exercise 14.20 with some observations deleted. (An entry of "999" represents a missing observation.

SIT			**PE**			**SC**		
Pre	**Post**	**Followup**	**Pre**	**Post**	**Followup**	**Pre**	**Post**	**Followup**
19	6	1	20	5	0	12	14	18
28	14	16	999	999	21	27	18	9
18	6	8	36	26	17	24	999	13
999	6	11	25	11	9	32	21	11
21	6	13	26	999	7	26	20	18
24	10	8	30	31	10	18	20	26
26	10	999	19	6	11	38	35	34
15	6	13	19	7	999	26	22	999
18	8	6	22	4	999	23	10	8
34	13	8	22	17	20	22	19	19
20	999	999	24	19	1	34	999	999
34	10	1	28	22	16	22	15	12
29	16	23	29	23	20	27	18	13
33	19	39	27	15	20	23	21	19
22	7	16	27	7	3	26	18	13

a) Analyze these data using a standard repeated measures analysis of variance.

b) How do your results differ from the results you found in Exercise 14.20?

14.23) Now analyze the data in Exercise 14.22 using a mixed models approach, an appropriate form for the covariance matrix. How do those results differ from the results you found in Exercise 14.22?

14.24) In the data file Stress.dat, available on the Web site, are data on the stress level reported by cancer patients and their spouses at two different times—shortly after the diagnosis and 3 months later. The data are also distinguished by the gender of the respondent. As usual, a "." indicates each missing data point. See description in Appendix: Computer Data Sets, p. XXX.

a) Use any statistical package to run a repeated-measures analysis of variance with Gender and Role (patient versus spouse) as between-subject variables and Time as the repeated measure.

b) Have the program print out cell means, and plot these means as an aid in interpretation.

c) There is a significant three-way interaction in this analysis. Interpret it along with the main effects.

14.25) Everitt reported data on a study of three treatments for anorexia in young girls. One treatment was cognitive behavior therapy, a second was a control condition with no therapy, and a third was a family therapy condition. The data follow.

Group	Pretest	Posttest	Gain
1	80.5	82.2	1.7
1	84.9	85.6	.7
1	81.5	81.4	-.1
1	82.6	81.9	-.7
1	79.9	76.4	-3.5
1	88.7	103.6	14.9
1	94.9	98.4	3.5
1	76.3	93.4	17.1
1	81.0	73.4	-7.6
1	80.5	82.1	1.6
1	85.0	96.7	11.7
1	89.2	95.3	6.1
1	81.3	82.4	1.1
1	76.5	72.5	-4.0
1	70.0	90.9	20.9
1	80.4	71.3	-9.1
1	83.3	85.4	2.1
1	83.0	81.6	-1.4
1	87.7	89.1	1.4
1	84.2	83.9	-.3
1	86.4	82.7	-3.7
1	76.5	75.7	-.8
1	80.2	82.6	2.4
1	87.8	100.4	12.6
1	83.3	85.2	1.9
1	79.7	83.6	3.9
1	84.5	84.6	.1
1	80.8	96.2	15.4
1	87.4	86.7	-.7
2	80.7	80.2	-.5
2	89.4	80.1	-9.3
2	91.8	86.4	-5.4
2	74.0	86.3	12.3
2	78.1	76.1	-2.0
2	88.3	78.1	-10.2
2	87.3	75.1	-12.2

Group	Pretest	Posttest	Gain
2	75.1	86.7	11.6
2	80.6	73.5	-7.1
2	78.4	84.6	6.2
2	77.6	77.4	-0.2
2	88.7	79.5	-9.2
2	81.3	89.6	8.3
2	78.1	81.4	3.3
2	70.5	81.8	11.3
2	77.3	77.3	0.0
2	85.2	84.2	-1.0
2	86.0	75.4	-10.6
2	84.1	79.5	-4.6

2	79.7	73.0	-6.7
2	85.5	88.3	2.8
2	84.4	84.7	0.3
2	79.6	81.4	1.8
2	77.5	81.2	3.7
2	72.3	88.2	15.9
2	89.0	78.8	-10.2
3	83.8	95.2	11.4
3	83.3	94.3	11.0
3	86.0	91.5	5.5
3	82.5	91.9	9.4
3	86.7	100.3	13.6
3	79.6	76.7	-2.9
3	76.9	76.8	-0.1
3	94.2	101.6	7.4
3	73.4	94.9	21.5
3	80.5	75.2	-5.3
3	81.6	77.8	-3.8
3	82.1	95.5	13.4
3	77.6	90.7	13.1
3	83.5	92.5	9.0
3	89.9	93.8	3.9
3	86.0	91.7	5.7
3	87.3	98.0	10.7

a) Run an analysis of variance on group differences in Gain scores.

b) Repeat the analysis, but this time use a repeated measures design where the repeated measures are Pretest and Posttest.

c) How does the answer to part (b) relate to the answer to part (a)?

d) Plot scatterplots of the relationship between Pretest and Posttest separately for each group. What do these plots show?

e) Run a test on the null hypothesis that the Gain for the Control is 0.00. What does this analysis tell you? Are you surprised?

f) Why would significant gains in the two experimental groups not be interpretable without the control group?

Discussion Questions

14.26) In Exercise 14.24 we ignored the fact that we have pairs of subjects from the same family.

a) What is wrong with doing this?

b) Under what conditions would it be acceptable to ignore this problem?

c) What alternative analyses would you suggest?

14.27) In Exercise 14.24 you probably noticed that many observations at Time 2 are missing. (This is partly because for many patients it had not yet been 3 months since the diagnosis.)

a) Compare the means at Time 1 for those subjects who did, and who did not, have data at Time 2.

b) If there are differences in (a), what would this suggest to you about the data?

Not Numbered In a study of behavior problems in children we asked 3 "judges" to rate each of 20 children on the level of aggressive behavior. These judges were the child's Parent, the child's Teacher, and the child him/herself (Self). The data follow.

Child	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Parent	10	12	14	8	16	21	10	15	18	6	22	14	19	22	11	14	18	25	22	7
Teacher	8	13	17	10	18	24	9	16	18	8	24	19	15	20	10	18	19	30	20	10
Self	12	17	16	15	24	24	13	17	21	13	29	23	16	20	15	17	21	25	25	14

These data are somewhat different from the data we saw in Section 14.10 because in that case the same people judged each child, whereas here the Parent and Self obviously change from child to child. We will ignore that for the moment and simply

act as if we could somehow have the same parent and the same "self" do all the ratings.

14.28 What is the reliability of this data set in terms of the intraclass correlation coefficient?

14.29 What do your calculations tell you about the sources of variability in this data set?

14.30 Suppose that you had no concern about the fact that one source systematically rates children higher or lower than another source. How might you evaluate reliability differently?

14.31 Under what conditions might you not be interested in differences among judges?

14.32 What do you think is the importance of the fact that the "parent" who supplies the parent rating changes from child to child?

14.33 Strayer, Drews, & Crouch (2006) (which we saw as a between-subjects design in Exercise 11.32) examined the effects of cell phone use on driving ability. They had 40 drivers drive while speaking on a cell phone, drive while at the legal limit for alcohol (0.08%), and drive under normal conditions. (The conditions were counterbalanced across drivers.) The data for this study are found at www.uvm.edu/~dhowell/methods/DataFiles/Ex14-34. Their hypothesis, based on the research of others, was that driving while speaking on a cell phone would have as much of an effect as driving while intoxicated. The dependent variable in this example is "braking reaction time." The data have exactly the same means and standard deviations as they found.

a) Run the analysis of variance for a repeated measures design.

b) Use the appropriate contrasts to compare the three conditions. Did the results support the experimenters' predictions?

[1] This assumption is overly stringent and will shortly be relaxed somewhat. It is nonetheless a sufficient assumption, and it is made often.
[2] Because I have rounded the means to three decimal places, there is rounding error in the answers. The answers given here have been based on more decimal places.
[3] Both SPSS and SAS continue to calculate the wrong value for the Huynh-Feldt epsilon.
[4] The authors used a logarithmic transformation here because the original data were very positively skewed. They took the log of $(X + 1)$ instead of X because log(0) is not defined.
[5] As in earlier tables of expected mean squares, we use the $\in\sigma^2$ to refer to the variance of random terms and θ^2 to refer to the variability of fixed terms. Subjects are always treated as random, whereas in this study the two main independent variables are fixed.
[6] For those who want to see the calculations, the corresponding pages from the previous edition can be found at www.uvm.edu/~dhowell/methods/whateverIcallit.html.
[7] In previous editions I used the MANOVA approach under SPSS/Univariate/Repeated measures as a way of avoiding assumptions of compound symmetry. This approach does not require compound symmetry, but it does require balanced designs. I have dropped it in favor of the mixed model precisely because the mixed model will handle missing data much better.
[8] The following is quick description of using the menu selections. Select **analysis/mixed/linear,** specify Subj for the Subjects box and Time for the Repeated box. Click **continue** and move to the next screen. Specify the dependent variable (dv) and the factors (Group and Time). Select **fixed** from the bottom of the box, highlight both Group and Time and click the **add** button, click **continue**. Now click on the **random** button and add Subj to the bottom box. Then click **paste** to make sure that you have syntax similar to what I gave above.