17

The Single-Factor Within-Subject Design: Further Topics

In the last chapter, we presented the techniques you need to conduct and interpret a single-factor within-subject analysis of variance. As we indicated, this design offers a substantial gain in power and convenience. However, these advantages, like all such benefits, come with certain costs. A within-subject design requires changes in the assumptions that underlie the tests, making it more likely that they will be violated. We discuss these issues and their correction in this chapter.

17.1 Advantages and Limitations

We start with a comparison of the between-subjects and within-subject designs. In brief, a study conducted with a within-subject design obtains more data from each subject than one conducted with a between-subjects design, and the analysis has a smaller error term. Set against these gains are the facts that repeated observations of a subject cannot be collected under constant conditions, and that any earlier observation has the potential to influence later ones. Moreover, the assumptions that underlie the analysis are more complex than those of the between-subjects designs.

Advantages of the Within-Subject Design

The three principal advantages of a within-subject design over a between-subjects design are more efficient use of subject resources, greater comparability of the conditions, and reduced error variance. The most obvious of these is the economy of the design. By taking several observations from each subject, much more data can be collected in a short period of time. The advantage is particularly great when each subject is "expensive," either because only a few are available or because a considerable amount of preparation or instruction is necessary before the study can begin. The efficiency of the within-subject designs certainly accounts for their wide use in behavioral research.

From a statistical point of view, a great advantage of the within-subject designs is its increased control of subject variability. The subjects that come into an experiment are not equivalent in ability, or on any other factor, for that matter. In a between-subjects design, they are randomly assigned to independent groups. Although random assignment eliminates the chance of *systematic* differences among the groups (other than those that arise from the treatments), *accidental* differences among groups will arise because different subject are assigned to each group. These chance differences are superimposed on whatever treatment effects are produced by the experimental manipulations. In the within-subject design, we select a single group of subjects and have them serve in every treatment condition, thus ensuring that comparable subject differences are present in each condition.

The third advantage of the within-subject design goes along with the second. By making the conditions more similar, the size of the error term used to test for differences among the treatments is reduced. Any study involves two types of random variability, one reflecting consistent differences among the subjects, the other reflecting variation from one observation to another. Among the second set are momentary changes in attention and motivation, and variation in the physical environment or the testing apparatus. Both types of variability affect the accidental differences among treatments in a between-subjects design, but only the second affects differences among treatments in a within-subject design. As you saw in the last chapter, the appropriate error term to compare with the within-subject treatment effect is the treatment-by-subject interaction. This variability is almost always less than the pooled within-group variability used to test the effect in a between-subjects design, which increases the test's power.

Limitations of Within-Subject Designs

The within-subject design has both statistical and nonstatistical limitations. The statistical problems mostly concern the sensitivity of the assumptions of the analysis. The scores produced by a single subject are more alike than are the scores produced by different subjects. In a statistical sense, this similarity means that the observations are not independent. The model on which the analysis is based must say something about this dependence, and these extra assumptions make the model more complex and more vulnerable to violation. In contrast, in a between-subjects design, the experimental procedure assures that the observations from different subjects are independent. Thus, there is a simplicity to a between-subjects design that is absent from its within-subject counterpart. We will discuss these issues in Sections 17.2 and 17.3.

The nonstatistical problems arise from the fact that the repeated observations must necessarily take place under somewhat different conditions, and some aspect of this difference, other than the treatment being investigated, can affect the scores. These differences do not affect between-subjects designs, because only a single score is recorded. We will refer to these differences as **incidental effects**—systematic differences that are incidental to the actual treatment manipulation. We will review them here and then discuss methods for controlling them in Sections 17.4 and 17.5.

Incidental differences among the treatments can arise for a variety of reasons. One general class concerns nonspecific changes in the subject, such as practice and fatigue. If a subject becomes tired or bored, performance will drop for the later observations, regardless of the treatment involved. Similarly, if a subject becomes better at the task or at following instructions, performance will improve. In an intermixed-treatment design, changes like these affect all treatments equally. They may increase the variability of the data, but they do not bias the results. In a successive-treatment design, however, the conditions are affected unequally. Practice effects help the conditions administered late in testing, while fatigue effects hurt them. Effects of this type should be considered whenever a successive-treatment study is planned.

A second form of incidental differences arises when different treatments must use different material. In an experiment on memory, for example, subjects must be given a different set of material to learn with each new treatment—they cannot relearn the same material over and over again. If one set of material is easier to remember than another, it will give an advantage to the treatment condition in which it is used. Effects of this type influence both successive-treatment and intermixed-treatment designs.

The other types of nonstatistical problems relate to the specific nature of the treatments. These can roughly be divided into three groups, each with different causes and cures: carryover effects, contrast effects, and context effects. A **carryover effect** occurs when a treatment has a transient effect that carries over to affect whatever condition is administered immediately after it. Consider a study that evaluates the effect of a drug by looking at behavior following doses of different sizes. A large dose may depress behavior, both immediately and for some period of time thereafter. If the tests are not spaced by a sufficiently long interval, the next test will be affected not only by the currently administered drug, but by the continued effect of the earlier drug. In contrast, a small dose (or particularly a placebo control) has little or no continuing effect. Carryover effects can be both physical and psychological. For example, Sheehe and Bross (1961) remark that the effectiveness of an analgesic agent in reducing the perception of pain is substantially reduced when it follows an ineffective agent, even when sufficient time has elapsed that no possibility of chemical mediation exists. In this case, the carryover is psychological, with the patients perhaps temporarily losing confidence in painkillers.

A **contrast effect** is a carryover effect that occurs when two treatments interact in a way that depends on both conditions. Suppose a researcher is studying the effects of giving praise, reproof, or no feedback during a learning task with second-grade children. The effect of no feedback, say, may be quite different when it follows the praise condition than when it follows the reproof condition. Another example is the reward study mentioned in the introduction to Part V (p. 348). The 5-cent reward for a successful outcome might be valued more if it is preceeded by a 1-cent reward than if it is preceeded by a 50-cent reward. Carryover effects of these sorts are possible only in a within-subject design, and where they are severe, the best approach may be to use a between-subjects design. Where they are more mild, it may be sufficient simply to ensure that the conditions are not always presented in the same order.

The very act of being measured in one treatment condition can change the subject, so that later observations are contaminated. With learned material, for example, it is impossible to test a subject on the same material twice, because the first test functions as an additional learning trial, and influences the later test. A subject who has been tested 10 minutes after learning, for example, is likely to remember more at 30 minutes than one who was not tested. There are more extreme forms of this phenomenon, generally known as **context effects**, in which a subject's behavior is influenced by the context provided by exposure to other conditions in an experiment. A surprise test can only be given once; after that subjects will expect additional tests and alter their behavior accordingly. Studies of incidental learning, in which subjects are tested for what they can remember of material presented without formal instructions to learn, have this problem. Once they have been tested the first time, subjects will expect additional tests and later learning will not be incidental. The most extreme form of these effects come from the physiological manipulations such as an operation, which obviously cannot be undone.

Consider a few other examples of potential context effects. If subjects are told that performance on a given task is a measure of intelligence (as a way of increasing their motivation), how will they view any future tasks where the set is changed? If subjects are told to employ one strategy when learning a set of material, will they be able to adopt a new one when the conditions are switched? If subjects are misled about what will happen in one treatment condition—a technique used in so-called deception experiments—will they believe what the researcher says about later treatments? If some of the experimental conditions are frightening or distasteful, how will subjects react when they are told that other conditions will be milder? For example, if they have been punished for an error with an electric shock, will they be unaffected by this experience when they are told they will not be shocked in another condition? These are all examples of situations that probably should not be studied in within-subject designs. Greenwald (1976) provides a useful discussion of these sorts of problems.

The discovery that an otherwise well-designed within-subject study is subject to serious carryover or context effects does not completely render the study useless. When each treatment has appeared first for some of the subjects, it is possible to analyze just the data from the first testing session. Performance at this point is completely uncontaminated by the effects of prior testing. Although this retreat to a between-subjects design loses the increased power and comparability of the groups associated with a within-subject design, it restores the interpretability of the results, which is obviously the most important consideration.

17.2 The Statistical Model

Like all statistical procedures, the analysis of the within-subject design is based on a set of assumptions about how the data are produced—the statistical model for a score Y_{ij}. Although the configuration of the data looks superficially like that of a between-subjects design, there are critical differences in the way the scores are represented. Here we develop the statistical model, and in the next section we discuss the effect of violating one of the key assumptions.

The difference between the models for the between-subjects and within-subject designs lies in the assumption of independence of the scores. The fact that several scores come from the same subject causes the scores in different conditions to be correlated—for example, a subject that produces a high score under one treatment is likely to have a high score under another. We can observe this dependence by calculating the correlation coefficients among the three conditions in the numerical example from the last chapter (Table 16.3). When we take the six scores for a_1 and correlate them with the matching six scores for a_2, we find $r_{12} = 0.938$. The correlations between the other two pairs of scores are also substantially greater than zero,

$r_{13} = 0.947$ and $r_{23} = 0.937$. The within-subject observations are not independent of each other, and the statistical model needs to accommodate these dependences.

Two different models have been applied to within-subject data. They differ in whether scores from a subject are treated as separate entities or are treated together. In the **univariate approach**, each score Y_{ij} is viewed as a separate random variable made up of systematic and random components, including a component specific to the subject. In the **multivariate approach**, all the scores from a single subject are treated as a single statistical entity. In Table 16.3, for example, the univariate approach treats the first subject's scores as $Y_{11} = 745$, $Y_{12} = 764$, and $Y_{13} = 774$, but the multivariate approach treats them as $\mathbf{Y}_1 = (745, 764, 774)$; we use the boldface letter to indicate that the observation is multivariate. The two representations have different sets of assumptions and lead to two different forms of the analysis of variance. The univariate approach leads to the analysis described in Chapter 16, and the multivariate approach leads to the **multivariate analysis of variance** or **MANOVA**. The multivariate analysis is the more flexible of the two, and it requires fewer assumptions about the data. However, as a result, when this flexibility is not required, it is less powerful than the univariate approach. We will emphasize the univariate approach for several reasons. First, its assumptions are often satisfied, or nearly so, with experimental data. Second, it lends itself to the type of multifactor experiments that are common in behavioral research. Both of these properties have made it popular with researchers. Finally, it is computationally easier than the multivariate model, making it a good place to develop one's understanding of the techniques.

Both models require the usual cluster of random-sampling assumptions. In particular, each subject's data must be independent of the data from every other subject, and the same distribution must apply to every subject. Note that the assumption of independence applies to the subjects, not to the individual scores, which will be correlated because of the consistency of subjects. As always, these assumptions are critical to the analysis. Without them, the population to which the inferences apply is unclear, and the results of the analysis are potentially biased or irrelevant. The two models also specify, in somewhat different forms, the assumption of a normal distribution, as we will amplify below.

The Univariate Model

One way to think of the univariate model for the within-subject analysis of variance is as a specialized form of the randomized-blocks design from Section 11.5. In that design, blocks of subjects were formed whose scores were expected to be relatively similar. Subjects within these blocks were then randomly assigned to the conditions, and the blocks were treated as a factor in the design. Because the variability associated with differences among blocks was systematically removed, the treatment effects were both better balanced and less variable than they would have been in a simple between-subjects design. In applying this model to a within-subject design, we treat each subject as a block. Like a block in the randomized-blocks design, the scores for a single subject in a within-subject design are more similar to each other than they are to the scores of the other subjects.

The parallel to a randomized-blocks design correctly suggests that a linear model similar to the two-factor model for that design can be used (Equation 11.14, p. 225).

There are some important differences, however, in the nature of any uncontrolled error. In a randomized-blocks design, the blocks are chosen systematically, while in a within-subject design, the subjects are chosen randomly. Moreover, although there are often several subjects in each condition within a block, subjects in the within-subject design provide only one score per condition. With these modifications, a score Y_{ij} is expressed by the equation

$$\mathsf{Y}_{ij} = \mu_T + \alpha_j + \mathsf{S}_i + (\mathsf{S}\alpha)_{ij} + \mathsf{E}_{ij}. \tag{17.1}$$

The grand mean μ_T and the treatment effect α_j are familiar. The remaining terms define the different sources of unsystematic variability:

1. *The overall ability of the subject,* S_i. Some subjects produce, on average, high scores and others low. The random variable S_i represents the deviation of these mean scores from the grand mean. It has a normal distribution with a mean of 0 and a variance of σ_S^2.
2. *The idiosyncratic response of the subject in a particular condition,*[1] $(\mathsf{S}\alpha)_{ij}$. Differences in skill, ability, or predilection make some subjects perform better in one condition, others in another. These effects constitute a treatment-by-subject interaction. Because the subjects are randomly selected, it is represented by a random variable. It has a normal distribution, with a mean of 0 and a variance of $\sigma_{A\times S}^2$.
3. *The variability of the individual observations,* E_{ij}. Even in the same condition, a particular subject would not produce the identical score each time the same treatment was administered. The uncertainty about this aspect of performance is represented by the random variable E_{ij}, which has a normal distribution with a mean of 0 and a variance of σ_{error}^2.

You can compare this model to the linear model for a one-way between-subjects design, $\mathsf{Y}_{ij} = \mu_T + \alpha_j + \mathsf{E}_{ij}$. Although the three sources of variability S_i, $(\mathsf{S}\alpha)_{ij}$, and E_{ij} are present there (see the discussion of experimental error in Chapter 2 on pp. 19–20), they are indistinguishable. As a result, the variabilities σ_S^2, $\sigma_{A\times S}^2$, and σ_{error}^2 of the within-subject design are lumped together as the single error variance σ_{error}^2 of the between-subjects design.

The effect of the three sources of unsystematic variability is to make the mean squares more complicated than they were in the between-subjects designs. Calculations (which we will not go through) show that

$$\left.\begin{aligned} \mathrm{E}(MS_A) &= \frac{n}{a-1}\sum \alpha_j^2 + \sigma_{A\times S}^2 + \sigma_{\text{error}}^2, \\ \mathrm{E}(MS_S) &= a\sigma_S^2 + \sigma_{\text{error}}^2, \\ \mathrm{E}(MS_{A\times S}) &= \sigma_{A\times S}^2 + \sigma_{\text{error}}^2. \end{aligned}\right\} \tag{17.2}$$

Let's look at the terms that make up these mean squares. The mean square for the treatment effect A, which is the term we want to test, is influenced by two of the three sources of error: the interaction of the treatment factor with subjects and the

[1] We have reversed the order of the two letters in this effect from that used in earlier editions of this book. This change keeps their order and that of the subscripts in alignment, with the first (S or i) referring to subjects and the second (α or j) to the treatment conditions.

variability of the individual observations. Under the null hypothesis, when all the treatment effects α_j are zero, $\mathrm{E}(MS_A)$ reduces to $\sigma^2_{A\times S} + \sigma^2_{\text{error}}$. The denominator of the F ratio must match this sum, and the $A\times S$ interaction does. It should be clear now why the error term for the A effect is the $A\times S$ interaction. The expected mean squares also show why there is no pure test of the S effect, as we mentioned on pages 353–354. The error variance σ^2_{error}, which would be the necessary term, never appears in isolation, so it cannot be independently estimated. An F test that used $MS_{A\times S}$ as an error term would be biased.[2]

The univariate model constrains the possibilities for the variances of the scores and the correlations among them. Calculations based on Equation 17.1 show that when it holds, two things happen. First, the variances of all the treatment conditions are identical. Second, the same thing happens to correlations between the scores; they too are identical. These conditions are often referred to as **homogeneity of variance** and **homogeneity of correlation**, respectively. When these restrictions hold, the data are said to show **compound symmetry**. We will talk more about the implications of these restrictions in Section 17.3.

The Multivariate Model

The alternative to the univariate representation is the multivariate model. This model treats all the scores from a subject as a single multipart random variable that contains the individual scores within it. The complete set of observations for subject s_i is the multivariate random variable

$$\mathsf{Y}_i = (\mathsf{Y}_{i1}, \mathsf{Y}_{i2}, \ldots, \mathsf{Y}_{ia}). \tag{17.3}$$

The model itself simply says that this random variable has what is known as a **multivariate normal distribution**, which is a generalization of the normal distribution that allows for the correlations among several variables. The parameters of this model, corresponding to the μ, α_j, etc. of the univariate model, are the parameters of the multivariate distribution. They are of three types: the means μ_j of the individual scores, the variances σ^2_j of these scores, and the correlations between the pairs of scores. The null hypothesis tested by the multivariate analysis of variance is the same as that tested by the univariate form, namely, that the μ_j are all identical.

The way the two models differ is in how the variability among the scores is expressed. We saw above that the univariate model imposes compound symmetry on the data. All the scores must have the same variance, and the correlations between any pair of scores must be the same. The multivariate model completely relaxes this requirement. It can accommodate any pattern of variances and correlations. Because of this flexibility, it applies to situations for which the univariate model is inappropriate. This robustness comes at a cost, however. The multivariate analysis, in effect, must estimate all those variances and correlations from the data, and this process reduces the amount of information that can be brought to bear on the differences among the means. As a result, when the assumptions of the univariate model hold, the multivariate tests have less power. Without going into the details, this reduction

[2] These quantities allow us to complete the equations for the effect size that we expressed by words in Equation 16.7. The total variability is $\sigma^2_A + \sigma^2_S + \sigma^2_{A\times S} + \sigma^2_{\text{error}}$, and the variability that affects $A\times S$ is just $\sigma^2_{A\times S} + \sigma^2_{\text{error}}$.

in power is reflected by the denominator degrees of freedom under the two models. For the example in Table 16.3, $df = 10$ for the univariate model and $df = 4$ for the multivariate model, leading to critical values for the respective F tests of $F_{.05}(2, 10) = 4.10$ and $F_{.05}(2, 4) = 6.94$.

The multivariate analysis of variance gives rise to several different test statistics, and the programs (which you will surely use to conduct this procedure) usually give several of them. They all give the same result for completely within-subject designs—those that have only within-subject factors. We will return to the different test statistics when we consider the mixed design.

There is one place where the two approach are the same. The test of a contrast under the multivariate model uses the procedure described in Section 16.2. The calculation of the contrast-by-subjects interaction as an error term for each contrast has the effect of matching the error to the observed variability of that contrast, which is exactly what the multivariate model does. We will discuss below some situations where the univariate approach can be used and some where the multivariate approach is better. You can disregard these distinctions when you are testing contrasts.

17.3 The Sphericity Assumption

We mentioned that the univariate model implies the twin conditions of homogeneity of variance and homogeneity of correlation, that is, of compound symmetry. These assumptions ensure that tests based on the univariate model are valid. Actually, because the hypothesis of no treatment effect concerns only *differences* between scores, a slightly weaker assumption is all that is needed. Compound symmetry need not hold for the scores themselves, but only for the differences between pairs of scores. This condition is referred to as **circularity** or **sphericity** (the difference between these terms need not concern us, and we will use the latter term). We will emphasize the somewhat more stringent assumption of compound symmetry, because it is expressed directly in terms of the data and easier to relate to an actual study.

To avoid violations of compound symmetry, the various observations should be of the same type and measured in a similar way. Of course, measures of completely different quantities, such a response time and the proportion of errors, should never be treated as a within-subject "factor," even when they are collected on the same subjects. However, even measures that seem superficially similar may create a problem. Consider a memory researcher who asks subjects to recall a story. The facts are divided into several categories (for example, those describing the characters, those describing the action, those that are incidental to the story, etc.), and the researcher wants to compare the proportion of each type that each subject recalls. The problem here is that these proportions, being based on different numbers of original facts and having different rates of recall, are also likely to have systematically different variances. The correlations among them may also vary. To take another example, suppose a researcher has measured the skills of a group of children using tests of verbal skills, arithmetic skills, and motor skills. Although all of these tests capture the idea of *skill*, they measure different concepts. The school-related measures of verbal and arithmetic skills are likely to be more highly correlated with each other than either is to the motor task. In both cases, a test based on the multivariate model should be used.

Time-series data are another source of violations of compound symmetry. Suppose the same measure is taken from the subjects at different times, such as on different trials in a learning experiment. With these data, adjacent scores are likely to be more highly correlated than are scores observed at widely spaced intervals. Even when the variances are the same, compound symmetry may fail because the correlations are heterogeneous. A similar situation arises in a study that measures changes during the course of therapy. Suppose a standardized measure of pathology is administered before treatment is started, at the midpoint of treatment, and after the treatment is finished, and these values are to be compared. The correlation between the scores measured before and after therapy is likely to be lower than the correlation between either of them and the middle scores. The danger of temporal heterogeneity is particularly high when many measurements have been taken.

There are tests for violations of sphericity or compound symmetry. The most widely used of these, a likelihood-ratio test statistic W developed by Mauchly (1940), is included in a number of computer programs. This statistic should *not* be significant for the analysis to proceed. For example, a computer analysis of the data in Table 16.3 gave $W = 0.758$, with a descriptive level of $p = 0.575$. Because this value is far from significant, we can proceed with the within-subject analysis of variance without concern. There are some important limitations to Mauchly's test, however. It has been criticized both for a lack of sensitivity to small violations (which can nevertheless affect the F test) and for positive biases when the data contain a disproportionate number of extreme scores. A test due to John (1971) is superior (see Cornell, Young, Seaman, & Kirk, 1992, for a comparison of several tests and references to earlier work, and Kirk, 1995, pp. 277–278, for examples of its use). It has not been widely implemented in the computer packages.

The tests of sphericity, like those of heterogeneity of variance in between-subjects designs, have their own assumptions. Their vulnerability to violations of these assumptions is not well understood, particularly those relating to failure of normality, but they are certainly different from those of the analysis of variance. Just as with homogeneity of variance in the between-subjects designs, you should be cautious about making decisions regarding the analysis procedure based only on apparent heterogeneity of the variances and correlations. Substantial violations force one to use the multivariate procedure, but in more ambiguous cases, the choice between approaches should be supported by a consideration of the measures themselves and an assessment of the plausibility of homogeneity and potential sources of heterogeneity. It is particularly confusing for readers of one's research to switch back and forth between different types of test statistic within the analysis of a single study or group of studies.

Dealing with Violations of Sphericity

When sphericity is violated, the omnibus tests based on the randomized-blocks model are biased positively. Tests using the critical value of F at, say, $\alpha = .05$ from Appendix A.1, may actually have a real, but unknown, significance level greater than .05. If we do not make some sort of adjustment, then we will be too likely to falsely reject the null hypothesis. There are four approaches we can take. Three of these attempt to salvage the omnibus test: measure the magnitude of violation of sphericity and adjust the critical value of F upward to accommodate it, use a conservative critical value based on the largest possible violation of heterogeneity, or turn to the multivariate approach. The fourth possibility is to forget about the omnibus test and emphasize tests of contrasts, which are immune to violations of sphericity.

The first way to eliminate the bias of the F test is to evaluate it against a larger critical value, obtained by reducing the degrees of freedom when entering Appendix A.1. Box (1954a) suggested using the values

$$df_{\text{num}} = \varepsilon(a-1) \quad \text{and} \quad df_{\text{denom}} = \varepsilon(a-1)(n-1), \tag{17.4}$$

where ε measures the extent to which sphericity is violated. When sphericity holds, $\varepsilon = 1$ and the degrees of freedom are those of the uncorrected test. When sphericity is violated, $\varepsilon < 1$, which reduces both df_{num} and df_{denom} and gives a larger critical value for F.

The problem now is to find a value for ε. Its true value depends on the actual variances and correlations in the population, which we do not know. There is no unambiguous way to estimate it. Among the possibilities are methods suggested by Geisser and Greenhouse (1958) and by Huynh and Feldt (1976), of which the latter has the greater power. Both values are calculated by many of the packaged programs.[3] Either method gives a larger critical value than the uncorrected test, which protects against an inflated Type I error rate.

Another approach is to pick the smallest value that ε can attain, whatever the variances and correlations may be, which happens to be $\varepsilon = 1/(a-1)$. Using this worst-case value, the observed F ratio is evaluated against the critical value obtained from Appendix A.1 with

$$df_{\text{num}} = \frac{a-1}{a-1} = 1 \quad \text{and} \quad df_{\text{denom}} = \frac{(a-1)(n-1)}{a-1} = n-1. \tag{17.5}$$

These values are those that would have been used had the study involved but two conditions, a fact that makes them easier to remember. Because this **conservative F test** was suggested by Geisser and Greenhouse (1958), it is often associated with their names, but you should distinguish it from the correction that uses their estimate of ε mentioned in the previous paragraph. The big advantages of the conservative test is that it is easy to use, requires no special tables, and can be applied even when access to the original data is no longer possible. Applied to the example of Table 16.3, the conservative degrees of freedom are

$$df_{\text{num}} = 1 \quad \text{and} \quad df_{\text{denom}} = n-1 = 6-1 = 5.$$

Looking in the F tables, we find that $F_{.05}(1,5) = 6.61$, which is larger than the unadjusted value of $F_{.05}(2,10) = 4.10$. Because the observed value of $F = 14.43$ exceeded the conservative criterion, we can reject the null hypothesis without worrying about violations of sphericity.

The difficulty with the conservative criterion is that it can give an ambiguous result. When the observed F falls between the unadjusted critical value and the conservative value—between 4.10 and 6.61 in our example—we can't tell whether to retain or reject the null hypothesis. The uncorrected criterion says "reject," and the conservative one says "retain." The researcher then has several alternatives, all of which require a computer. One possibility is to use one or the other of the specific values of ε to adjust

[3] Kirk (1995) and Myers and Well (1991) give examples of the calculations.

the degrees of freedom for the standard F ratio. Another possibility is to use a test based on the multivariate approach. It does not require the assumption of sphericity, so its results apply regardless of the pattern of variances and correlations in the data. A researcher should choose one of these approaches, not switch between them on an ad-hoc basis. On the whole, we favor the multivariate approach. It cleanly avoids the need to assume anything about sphericity, and it is based on a straightforward model for the data. It is currently more frequently used than the other possibilities. Any of these alternatives are less powerful than the ordinary F test when sphericity holds, so they should not be used unless necessary.[4]

In summary, we suggest that you start by looking at whatever indications of sphericity failures you have—a consideration of the design and the type of measures you are using and any statistics such as Mauchly's test. If you do not see difficulties, continue with the univariate test. If you do find cause to worry, try the conservative test. If the null hypothesis cannot be rejected using the standard criterion or can be rejected using the conservative criterion, you have your conclusion. If it falls in between, either use the multivariate test or focus on single-*df* hypotheses.

17.4 Incidental Effects

A subject's scores in any within-subject design are necessarily obtained under different conditions. Although, these differences are what distinguish the conditions that are being studied, they are affected by aspects of the experimental context that are incidental to the question under investigation. In a study using a successive-treatment design, the order in which the conditions are administered is the most obvious of these differences. One observation is made first, another second, and so forth. Unless the point of the study *is* the sequence of observations, as it would be if they were successive learning trials or patient status before, during, and after therapy, the order in which the treatments are given is incidental to the purposes of the study. Without taking this incidental factor into consideration, the experimenter would have to worry about whether performance on the later tests was improved because the early tests let the subject practice the task or was reduced because the subject became tired or bored with the study.

Other incidental aspects of the study apply to both successive-treatment and intermixed-treatment designs. The particular materials used in the task often must be varied in order to accommodate the repeated measurements. Consider a study in which subjects try to learn as much as possible about a briefly seen picture while performing some unrelated task, such as copying down the lyrics of a song or taking dictation over a telephone. Each subject serves in all conditions. The primary point of the study is to see which interfering activities makes the picture harder to remember. However, each time the subject is tested, a different picture must be used, and some of these pictures will be intrinsically easier to remember than the others. Always assigning the same picture to a given treatment condition confounds picture differences

[4] Their actual power depends on how sphericity is violated, which is unknown. There is some evidence that the various approaches have similar power under many natural ways that sphericity is violated with real data (Rasmussen, Heumann, Heumann, & Botzum, 1989).

with treatment effects. Like the order in which the conditions are administered, the differences among the pictures constitute an incidental aspect of the study.[5]

Factors such as the position in the testing sequence or the type of material are examples of the nuisance variables described in Chapter 1. When such a variable becomes an explicit factor in the design, we will refer to it as either a **nuisance factor** or an **incidental factor**. In a well-designed study the experimental conditions are not systematically affected by, or confounded with, such factors. If condition a_1 were always administered first, a_2 second, and so forth, we could never tell whether any differences in performance between them were due to the treatment or to practice or fatigue effects. Similarly, if a particular picture were always used for condition a_1, another always for condition a_2, and so on, we could never tell whether better performance in one condition was due to the treatment being studied or the use of an easier picture.

The biases that arise when the treatments are confounded with incidental aspects of the study, such as the order of testing or the materials, can be avoided by breaking up any consistent relationship between them. There are two ways to do this. In **randomization**, the relationship between the treatments and the incidental aspects of the study is chosen randomly; in **counterbalancing**, it is constructed in a way that systematically balances the incidental effects across the study. Each approach has advantages and disadvantages.[6]

Randomization

The randomization procedures are the easiest to apply. Little advanced planning is necessary. When each subject comes into the experiment, the order in which the conditions are administered is chosen randomly from among all possible orders, and any different types of material are randomly assigned to the conditions. The random orders break up any systematic relationship between the incidental aspects of the procedure and the treatment conditions. Randomization is particularly effective when several incidental factors must be accommodated, and the type of systematic counterbalancing that we will discuss next is prohibitively complex. Another advantage is that randomization does not require any special analysis procedures—the analysis we described in the last chapter applies without alteration.

Convenient as it is, randomization has two disadvantages. First, it cannot assure that the incidental factor is completely balanced across treatments. Just as it is unlikely that the random assignment of subjects in a between-subjects design results in perfectly equated groups, so the random assignment of materials or orders is unlikely to perfectly balance them over the treatment conditions. Second, the random variability of the incidental factors is absorbed in the error term $MS_{A\times S}$. Because the size of this term determines the power of the test, any procedure that can reduce it is helpful.

[5] We will discuss other ways in which the selection of material affects the statistical analysis in Chapter 24.

[6] The assignment of subjects to conditions in a between-subjects design may also be random or systematic. In a completely randomized design, subjects are assigned randomly; in a randomized-blocks design (Section 11.5), they are systematically grouped into blocks to create more homogeneous groups and to reduce the error term. Counterbalancing does the equivalent thing with nuisance variables.

Counterbalancing and the Latin Square

The alternative to randomization is counterbalancing: assigning an incidental factor systematically so each level occurs equally often with each treatment condition. In this way, any effects of the incidental factor apply equally to every condition. A counterbalanced design also allows the variablity of the incidental effect to be extracted as a sum of squares and eliminated from the error term. The result is to increase both the accuracy and the power of the tests. The downside of counterbalancing is the additional complexity introduced into the design—in the planning, in conducting the study itself, and in the analysis of the data.

The different methods of counterbalancing, particularly in the larger designs, can become very complicated. We will discuss only the simplest designs. Our discussion, we hope, will make the issues clear and help you deal with many practical situations. In more complex studies, particularly those with several nuisance factors, you will need to find a more comprehensive treatment or seek expert advice.

Suppose we are planning a study with four conditions that will be administered one after another in a successive-treatment design. The possibility of practice effects makes it unsatisfactory to administer the treatment conditions in the same order to all subjects—for example, condition a_1 first, a_2 second, and so forth. If we did, condition a_1 would have fresh but unpracticed subjects, and the later conditions would be affected by increasing amounts of fatigue and practice. To avoid this confounding, we can adopt the counterbalanced experimental plan in Table 17.1. The table on the left displays an idealized pattern of scores from four subjects tested on the four conditions. The columns represent the testing order, beginning with the first test, which we will denote p_1, through the fourth (p_4). Subject s_1 receives the treatments in the order a_1, a_2, a_3, a_4. For subject s_2, the order is changed to a_2, a_4, a_1, and a_3, and the other two subjects receive still different orders, as indicated in the table. Look at what this arrangement has done to the relationship between the treatments and the positions in which they were given. Treatment a_1 appears once in the first position, once in the second, once in the third, and once in the fourth. The same holds for the other three levels of factor A. No condition receives any advantage or disadvantage by appearing more often towards the beginning or the end of the testing sequence. Any practice or fatigue effects are spread evenly over the four treatment conditions.

The arrangement of the conditions in Table 17.1 is known as a **Latin square**. The name is derived from the fact that the pattern of conditions within the square is traditionally denoted by Latin letters—if we replace conditions a_1 to a_4 by the letters A, B, C, and D, respectively, the arrangement is

A	B	C	D
B	D	A	C
C	A	D	B
D	C	B	A

The key feature of the Latin square arrangement is that every letter appears exactly once in each row and each column. It is the basic tool that an experimenter uses to set up a design in which an incidental factor is systematically counterbalanced over the treatment conditions.

Table 17.1: An idealized pattern of scores from four subjects tested in a Latin square design. The left panel shows the data arranged by the order of testing; the right panel shows it arranged by condition.

	Testing position (P)				Treatment condition (A)			
Subject	p_1	p_2	p_3	p_4	a_1	a_2	a_3	a_4
s_1	a_1 4	a_2 11	a_3 8	a_4 14	p_1 4	p_2 11	p_3 8	p_4 14
s_2	a_2 7	a_4 10	a_1 11	a_3 9	p_3 11	p_1 7	p_4 9	p_2 10
s_3	a_3 1	a_1 8	a_4 13	a_2 15	p_2 8	p_4 15	p_1 1	p_3 13
s_4	a_4 6	a_3 5	a_2 14	a_1 12	p_4 12	p_3 14	p_2 5	p_1 6
Mean	4.50	8.50	11.50	12.50	8.75	11.75	5.75	10.75

On the right-hand side of Table 17.1, the data are rearranged so that the columns correspond to the treatment conditions, and each cell is labeled by the position in which that treatment was given. Each position appears once for each treatment, so the design is still a Latin square. At the bottom of both tables we have calculated the means—on the left, the means for each position; on the right those for each treatment condition. Main effects of both factors appear to be present. There is a practice effect with the means increasing with testing position from $\bar{Y}_{P_1} = 4.50$ to $\bar{Y}_{P_4} = 12.50$ and a treatment effect with means ranging from $\bar{Y}_{A_3} = 5.75$ to $\bar{Y}_{A_2} = 11.75$. Using the Latin square design allows each effect to be measured independently of the other.

We have plotted the scores from Table 17.1 as a function of condition and of testing position in Figure 17.1. We constructed the figure by extracting the four scores for each treatment condition (the columns in the right-hand part of the table), plotting each set according to testing position, and finally connecting each set of four points. This graph emphasizes the treatments and the testing positions but ignores the fact that the points on any of the lines were obtained from different subjects. Inspection of the figure reveals a marked practice effect, but one that is exactly the same for each treatment condition (this is why we referred to the pattern as "idealized" above); in addition, the clear separation of the curves suggests the presence of a treatment effect that is the same at each testing position. We know from the marginal means in Table 17.1 that there is a steady improvement in performance with successive testing positions and that condition a_3 is the worst and a_2 is the best, with the other two conditions in the middle.

This example shows how the failure to counterbalance would have corrupted the results. Look at the difference between the two extreme conditions, as revealed by the treatment column means in Table 17.1, $\bar{Y}_{A_2} - \bar{Y}_{A_3} = 11.75 - 5.75 = 6.00$. Suppose all the conditions had been tested in the order a_3, a_1, a_4, a_2, the same as the order given to subject s_3. The difference between these two conditions for this particular subject is the largest of all four subjects $(15-1 = 14)$. With this testing order, then, the practice

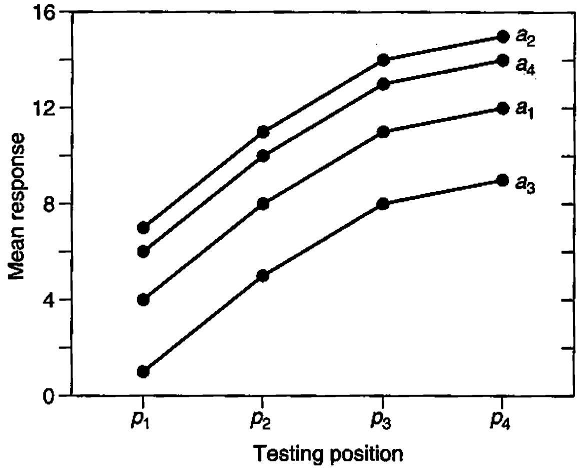

Figure 17.1: The scores in Table 17.1 plotted to show the treatment and practice effects.

effect mixes with the treatment effect in a way that exaggerates this difference. On the other hand, suppose the reverse order a_2, a_4, a_1, a_3 had been used, as it was for subject s_2. With this order, the practice effects work against the treatment effects—the difference between a_2 and a_3 is actually in the opposite direction ($7-9=-2$). No single testing order gives an uncontaminated measure of the treatment effect. Only by averaging over all the testing orders do we get a faithful picture of the effect.