CHAPTER 3

Designs That Incorporate Extraneous (Blocking) Factors

Frequently, experiments are run in the presence of extraneous factors that can affect the value of the response variable. Unless the effect of these factors is accounted for, misleading results may occur. For example, if suppliers of a particular raw material are to be compared in terms of the effect on a measure of roundness, a supplier effect could be confounded with an operator effect if different process operators take part in the experiment.

If the effect of one or more such extraneous factors is anticipated, a design can be constructed that will isolate such factors and separate them from the error term so that they will not affect the outcome of the computations and testing for the factor(s) of interest.

In this chapter we assume that there is one factor of interest and one or more blocking factors, although there is no reason why two or more factors could not be used with the design that is given in the next section, for example.

3.1 RANDOMIZED BLOCK DESIGN

Assume that the experiment to compare the suppliers must be run in one day so as to minimize the disruption on the regular production. Assume further that there is a morning/early afternoon shift, a late afternoon/evening shift, and a night shift. Assume that a different operator will be on duty for each shift at a critical part of the process while the experimentation is performed. If there were three suppliers and the raw material from the first supplier were used during the first shift, the raw material from the second supplier used during the second shift, and the raw material from the third supplier used during the third shift, the supplier effect and a possible operator effect would be completely confounded.

Modern Experimental Design By Thomas P. Ryan

The possible effect of the operator necessitates the use of a design with a *blocking factor*. If this is the only extraneous factor with which the experimenters are concerned, a *randomized complete block design* could be used. The word "complete" refers to the fact that every level of a factor (or every combination of factor levels if there is more than one factor) appears in each block, which are of equal size. We will occasionally refer to this as an RCB design for short. (Incomplete block designs are covered in Section 3.2.)

With this design, the extraneous factor is the blocking factor, and since there are three operators, there are three blocks. The general idea is to have experimental units that are more homogeneous within blocks than between blocks. Assume that there are also three suppliers. The number of levels of the factor of interest does not have to be the same as the number of blocks, however. Nine production units would thus be necessary for this experiment, and these production units would be the experimental units for the experiment. The raw material from the three suppliers would be assigned at random to the units within each block, and hence the name *randomized* block design. The design layout might appear as follows, with A, B, and C denoting the three suppliers.

	Blocks	
1	2	3
A	C	B
C	A	A
B	B	C

A sum of squares for blocks would be computed using the block totals, analogous to the way that the treatment (factor) sum of squares was shown to be computed in Section 2.1.3.3, as would the sum of squares for the suppliers since that is the treatment effect in this case.

Before proceeding further with this example, there are some questions that should be raised, and the assumptions for the design must also be considered. We can write the (typical) model for the design, analogous to the model for the completely randomized design in Eq. (2.1), as

$$Y_{ij} = \mu + A_i + B_j + \epsilon_{ij} \qquad i = 1, 2, \ldots, k \quad j = 1, 2, \ldots, t \tag{3.1}$$

As in Eq. (2.1), A_i is the effect of the ith treatment and B_j is the effect of the jth block, with Y_{ij} denoting the observation on the ith treatment in the jth block and ϵ_{ij} the corresponding error term, which is assumed to have a normal distribution with a mean of zero and a constant variance of σ^2 for each i, j combination.

3.1.1 Assumption

One obvious difference between Eq. (2.1) and Eq. (3.1) is that there is no true experimental error term, since ϵ_{ij} is the error term for the ith treatment and the jth block. That is, there cannot be a true experimental error term unless the (i, j)th combination

is repeated. (That *could* be done, but the resultant design would not be a conventional RCB design.) Instead, there is the assumption that there is no interaction between the treatment effect and the block effect. (Interaction is illustrated and discussed in detail in Section 4.2.) Simply stated, the assumption means that if we construct a scatterplot of blocks against treatments and appropriately connect the points to form lines that represent the response values for each block, the lines should be close to parallel. The crossing of lines, especially at sharp angles, would suggest that the no interaction assumption may not be valid. When this is the case, a formal analysis is not possible but the treatments might be ranked and compared within each block in an effort to extract some information from the data.

It is also generally assumed that the factor of interest is a fixed factor, whereas blocks are usually random (see Section 1.6.2.2 for a discussion of fixed and random factors). This means that the RCB model is a mixed model; that is, one factor is fixed and the other is random. Blocks might be fixed in some applications, however. Giesbrecht and Gumpertz (2004) have a lengthy discussion and some illustrations of blocks being random versus blocks being fixed. Somewhat similarly, even though the factor of interest is always assumed to be fixed, this does not have to be the case. For example, perhaps a large company is interested in the comparative performance of its machine operators, but doesn't want to test all of them. So a sample is selected and since the performance of each worker is suspected to depend on the age of the machine that is used, the workers are rotated (randomly assigned to) among the machines, which serve as the blocking factor.

If interaction between blocks and treatments is anticipated, either (a) the design should not be used, or (b) replication, as described above, should be used. Although it is good to be able to separate interaction from error if the former is feared and to be able to test for interaction, more than slight interaction will complicate the analysis, just as it complicates the analysis when designs presented in forthcoming chapters are used (see, e.g., Section 4.2.1). The main problem with a significant interaction is that the factor of interest (assuming a single factor) is tested against the error in the Analysis of Variance (ANOVA) table, with the error term representing both pure error and interaction, if the latter exists. A real interaction will thus inflate the error sum of squares, with the consequence that a significant factor effect may be declared not significant. As stated previously, the treatments in each block could be replicated if there were a concern about possible interaction, with the interaction then separated from error and tested against it. Again, however, this is not the way an RCB design is defined, but Exercise 3.26 contains an ANOVA table taken from the literature that has the two sources separated.

Relative to this experiment, the assumption means that, as a bare minimum, the ranking of the three suppliers and the separation between them (in terms of, say, closeness to a target roundness measure) are the same for each block (operator). That may or may not be a reasonable assumption. It would not be reasonable if one of the raw materials contained an ingredient to which one of the operators was slightly allergic, but the operator did not have such a problem with the other two raw materials, and the other two operators did not have a problem with any of the three raw materials.

If such a problem were to occur during the experiment, there would then be a true interaction, which should obviously not be used as the error term. In general,

interaction terms should never be used as a substitute for the experimental error term if this can be avoided. If it is economically and practically feasible to use multiple observations for each (i, j) combination, then this should be done.

Another factor to consider that will usually be more important is the number of observations needed to detect an effect of a given size, using an appropriate approach. Since n is fixed in the usual (unreplicated) randomized block design as it is determined by the number of blocks, Eq. (1.3) could be used (if at least a rough estimate of σ is available) to see if n is large enough to detect an effect of at least a desired size. Generally n will not be large enough unless at least a moderate number of blocks is used. The number of blocks that can be used will be determined by the design scenario, and cannot be freely chosen. For example, for the experiment being discussed, the number of blocks is fixed at 3, since that is the number of operators that are involved.

Thus, the usual unreplicated randomized block design should generally be eschewed in favor of a replicated design, with the number of replicates determined appropriately. The current example can be used to support that recommendation. Even though Eq. (1.1) is flawed, we will, because of its simplicity, still use it as a starting point in our analysis of the current problem and see how it performs relative to other methods. With three blocks and three suppliers, from Eq. (1.1) we have

$$\begin{aligned}\Delta &= \frac{4r\sigma}{\sqrt{n}} \\ &= \frac{4(3)}{\sqrt{9}}\sigma \\ &= 4\sigma\end{aligned}$$

remembering that this is (purportedly) for a power of .90 and a significance level of .05. By comparison, if we use Russ Lenth's sample size calculator (http://www.stat.uiowa.edu/∼rlenth/Power/index.html) and assume the use of Scheffé's method for comparing means, we find that the multiplier is 4.59, which does not differ greatly from the result obtained using Eq. (1.3), part of which is due to the difference in tests assumed.

This will be much too large a multiple of σ in a typical application; it would be necessary to use 12 blocks (four replicates) just to bring Δ down to 2σ, using both Eq. (1.1) and Lenth's applet. Thus, it should be apparent that there will erroneously be a failure to detect a significant difference in the means of a factor in many if not most applications of randomized block designs because of the number of blocks that are typically used.

A somewhat similar message is given by Dean and Voss (1999). They discuss a cotton-spinning experiment described by Peake (1953) that involved blocking. There were two treatment factors, "flyer" and "twist," with two levels of the first factor and four levels of the second factor. Two of the eight combinations were not observed and Dean and Voss (1999) subsequently converted this to, in essence, a one-factor design with six levels for the purpose of their analysis. Each experimental unit was

the production of a full set of bobbins on a single machine with a single operator, and the experimenters decided to use an RCB design with each block representing the condition of a single machine, a single operator, and a single week. Of course, if there were, say, an operator effect and a machine effect, they would be confounded with this blocking scheme, but that would be a minor problem as the intent was to compare the levels of the factor(s) free of extraneous effects.

The experimenters wanted to be able to detect a true difference of at least 2 (breaks per 100 pounds of material).

The block size was chosen to be 6, since this is the number of observations that could be made on a single machine in a single week, and of course was also equal to the number of treatment combinations. The experimenters decided to analyze the data after the first 13 blocks had been run, which of course took 13 weeks.

Dean and Voss (1999) addressed the question of how many blocks should have been used, and we will also address this question, but will use a different approach to arrive at the answer. With the assumption of $\sigma^2 = 7$, their approach was based on Scheffé's simultaneous confidence intervals and required a small amount of trial and error to arrive at 40 blocks, so that the experiment would require 40 weeks to run. Using Lenth's sample size calculator (http://www.stat.uiowa.edu/~rlenth/Power/index.html) and also using the Scheffé method, we arrive at 75 blocks if we are satisfied with a power of .90. (Power was not specified directly in the example given by Dean and Voss (1999); we should also note that the familywise error rate is what is controlled at the selected value (e.g., .05) when Lenth's applet is used.) These differences notwithstanding, it is apparent that a very large number of blocks is required.

Dean and Voss (1999, p. 307) showed that after 13 runs, the minimum significant difference in the means that could be detected was 3.57—almost double the desired value. Since only six observations could be made in a week, the block size could not be increased. The point to be made here is that a very large number of blocks was needed to accomplish the experimenter's objective, and this number far exceeded the number of levels of the factor. Thus, a simple textbook layout of a randomized block design with the number of blocks equal to or at least approximately equal to the number of levels of the factor would have been totally inadequate.

3.1.2 Blocking an Out-of-Control Process

It may also be necessary to block because of a process being out of control, as discussed in Section 1.7. Box, Bisgaard, and Fung (1990) recommended that processes be brought into as good a state of process control as possible and then blocking be used. This would not create a randomized block design, however, because the blocking could not be performed ahead of time, because times when processes go out of control and then stabilize cannot be forecast. (If one knew well ahead of time when a process would likely go out of control, then such events could likely be prevented. That is, the factors that cause an out-of-control process would be known, so there would be no point in conducting an experiment for the purpose of identifying them.) Instead, a practical way to view this use of blocking would be to form blocks posts-experimentation and hope that there is balance within each block. It isn't easy to

approximate the point at which a process has gone out of control, although some methods have been proposed for doing this.

3.1.3 Efficiency of a Randomized Block Design

The efficiency of an RCB design relative to a completely randomized design (CRD) has long been given in experimental design books. Efficiency figures in general are given as the ratio of two variances and in this case the ratio is the variance of a comparison of treatment levels without blocking divided by that variance when blocking is used. The efficiency expression is

$$\frac{(t-1)\text{MS}_{\text{blocks}} + t(k-1)\text{MS}_{\text{error}}}{(tk-1)\text{MS}_{\text{error}}}$$

If we write the fraction equivalently as

$$\frac{(t-1)\text{MS}_{\text{blocks}} + (tk-1)\text{MS}_{\text{error}} - (t-1)\text{MS}_{\text{error}}}{(tk-1)\text{MS}_{\text{error}}}$$

we can see that the ratio exceeds 1 if $\text{MS}_{\text{blocks}} > \text{MS}_{\text{error}}$. Since the F-test for testing whether there is a block effect is $F = \text{MS}_{\text{blocks}}/\text{MS}_{\text{error}}$, this means that the F-statistic will have to exceed 1.0, something that will frequently happen due to chance even when the blocking is not very effective.

There is a price that is paid if the blocking is ineffective, however, as the degrees of freedom for error is reduced when an RCB design is used. Specifically, with the CRD the error degrees of freedom is $n - k$, as shown in Section 2.1.3.3, whereas in an RCB design the error degrees of freedom is $n - k - (t - 1)$, with $n = kt$. Thus, the degrees of freedom for the RCB is smaller than for the CRD with the difference being the number of blocks minus one. When n is small, this reduction in the error degrees of freedom could cause a considerable loss of power to detect mean differences when blocking is not effective. Consequently, blocking should not be used injudiciously.

The main consideration, however, should be avoiding the wrong conclusion that could result from having a significant effect erroneously included in the error term. As with modeling in general, the emphasis should be on using a good model, and the randomized block model will often be the appropriate one.

3.1.4 Example

We consider the following experiment, described by Natrella (1963). Conversion gain, the ratio of available current-noise power to applied direct current power expressed in decibel units, is measured in six test sets, with four resistors used with each test set. The former served as blocks and the latter were the treatments. The data are in Table 3.1 with blocks and test sets numbered with consecutive integers, rather than using the identification numbers given by Natrella (1963).

TABLE 3.1 Dataset from Natrella (1963)

	Test Set					
Resistor	1	2	3	4	5	6
1	138.0	141.6	137.5	141.8	138.6	139.6
2	152.2	152.2	152.1	152.2	152.0	152.8
3	153.6	154.0	153.8	153.6	153.2	153.6
4	141.4	141.5	142.6	142.2	141.1	141.9

The first thing we can observe is that the blocking has been beneficial, since the conversion gain readings within each block are similar, but are dissimilar between blocks. We should plot blocks against treatments to check the interaction assumption, and the plot is given in Figure 3.1.

The configuration of points for the first resistor is curious and should perhaps lead to an investigation of the data for that resistor since this pattern does not exist for any of the other resistors. The overall configuration, however, is certainly acceptable as there is no strong indication of interaction.

The ANOVA table is given below. Computationally, the SS_{resistor} (i.e., SS_{blocks}) is computed in the same general way as $SS_{\text{test sets}}$ is computed, that is, as $\sum_{i=1}^{4} B_i^2 - (\text{GT})^2/24$, with B_i denoting the total of the ith block and GT denoting the

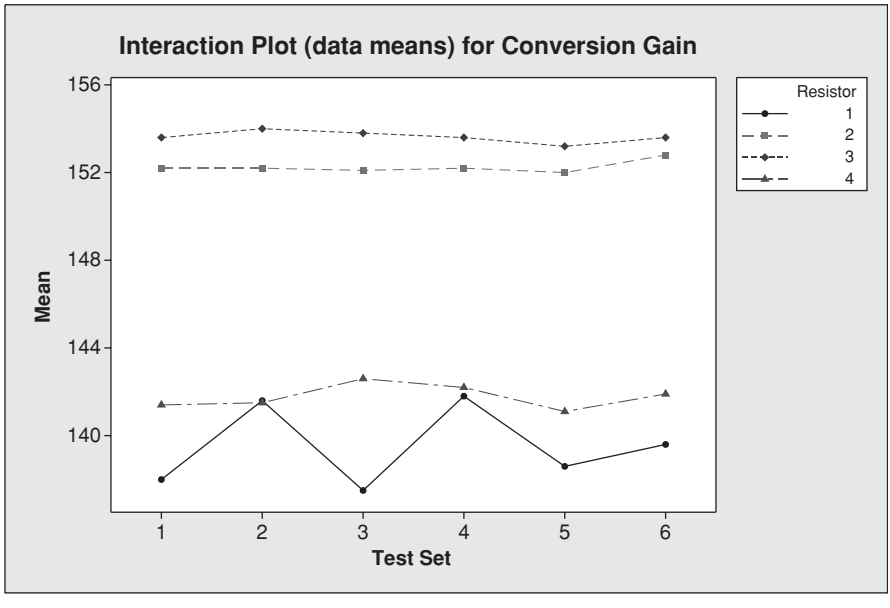

Figure 3.1 Interaction plot for Natrella data.

total of all the observations. The other computations are the same as those for a CRD.

Analysis of Variance for Conversion Gain

Source	DF	SS	MS	F	P
Resistor	3	927.66	309.22	344.40	0.000
Test Set	5	5.60	1.12	1.25	0.336
Error	15	13.47	0.90		
Total	23	946.73			

S= 0.947555 R-Sq = 98.58% R-Sq (adj) = 97.82%

We can see that the blocking was clearly effective, as evidenced by the large F-value for blocks (resistors) and by the fact that almost all of the variability in the conversion gain values is explained by blocks. The treatment factor (test set) is not close to being significant, but is closer than it would have been if blocking had not been used.

Curiously, Natrella (1963) stated, "We are interested in possible differences among treatments (test sets) and blocks (resistors)." Normally the only interest in blocks is to see if the blocking was effective, with there being only one factor of interest, the treatment effect, when an RCB design is used.

This raises the question of how the experiment should have been conducted if there had been two factors of interest. Designs for two or more factors are discussed in Chapter 4, but we will remark here that the randomization would have been different.

As stated previously, blocks are generally random in an RCB design, although fixed blocks are also used. In this application it may be appropriate to regard the blocks (resistors) as random, although that isn't clear. The statistical analysis is the same for both cases; only the inference is different, with the inference extending to a population of resistors if resistor is a random factor and to only the resistors used in the study if resistor is a fixed factor.

3.1.4.1 Critique

This is an example of an experiment for which replication was surely necessary, because the hypothesis of equality of six population means was tested with only four observations used to compute each sample average. Although there are enough degrees of freedom for estimating the error variance, the power might not be very good for detecting a mean difference that would be of practical significance. Whether or not this experiment should have been replicated by using either more blocks or multiple observations within blocks depends on the smallest difference in the treatment means that the experimenters wanted to detect, and there is no mention of this.

Surely the failure to recognize the need for adequate replication and the failure to determine how much replication is needed in a given experiment are among the major misuses of experimental designs.

If the hypothesis of equal treatment effects had been rejected, there would have been a need to determine which treatment levels differ. As with a CRD, the user can choose between a multiple comparison procedure, the sliding reference distribution approach mentioned in Section 2.1.3.1, or Analysis of Means (ANOM).

3.1.5 ANOM

As stated by Nelson (1993), ANOM can be used with any complete design. The only additional assumptions beyond those for ANOVA are that the effect estimates must be equicorrelated with correlation $-1/(k-1)$ for k means to be compared and the factor(s) must be fixed. Clearly the equicorrelation assumption is met for an RCB design since the effect estimators for the nonblocking factor have the same equicorrelation structure as in a CRD.

Examples of the ANOM used in conjunction with an RCB design are scarce in the literature; one example is given by Nelson, Coffin, and Copeland (2003, p. 330), and another example is given by Nelson, Wludyka, and Copeland (2005, p. 95). The use of ANOM with an RCB design presents no real complications or new considerations, however, as the means are plotted as in a CRD, with decision lines obtained from Eq. (2.2). Of course, blocks are considered to be random with an RCB design, so ANOM could not be applied to the blocks unless the blocks are fixed. Block analysis is generally not of any great interest, however. We would simply like to see evidence that the blocking has been beneficial after the analysis using blocks has been performed.

There are complications, however, in trying to use software to construct ANOM displays for this type of design. For example, MINITAB cannot be used to construct an ANOM display for an RCB design because it treats the data as having come from a two-factor design with interaction and there is no option for suppressing the two-factor interaction (which of course is assumed to not exist with the RCB design) and using that interaction as the error term.

Therefore, a MINITAB macro would have to be written (essentially starting from scratch) and programming would also have to be performed if SAS were used to construct an ANOM display.

Although not a suitable substitute for a graph, the test set means can be simply compared against the decision lines using Eq. (2.4), although that is hardly necessary here since the ANOVA showed the test set factor to be well short of significance. Nevertheless, we will illustrate the computations:

$$\begin{aligned}\bar{\bar{x}} \pm h_{\alpha,k,\nu}\, s\sqrt{(k-1)/(kn)} &= 146.8 \pm h_{.05,6,15}\sqrt{0.90}\sqrt{5/(6*4)} \\ &= 146.8 \pm 2.97(0.433) \\ &= (145.514,\ 148.086)\end{aligned}$$

The test set means are 146.30, 147.33, 146.50, 147.45, 146.23, and 146.98, all of which are well within the decision limits, as expected.

3.2 INCOMPLETE BLOCK DESIGNS

There will usually be physical constraints that will make it impossible to use all of the treatments in each block. For example, if days are blocks and different treatments are run through an industrial furnace in a manufacturing experiment, the furnace may not be capable of handling all the different treatments in one day. Although we usually associate physical and industrial experiments with the use of incomplete block designs, the designs are actually used in a wide variety of applications, including education (see, e.g., van der Linden, Veldkamp, and Carlson, 2004). Incomplete block designs are also often used in forestry, specifically in mixedwood and silviculture systems studies, for which it can be difficult to find blocks that are large enough to accommodate all the treatments. Incomplete block designs are also used in diallel cross experiments (Singh and Hinkelmann, 1999) because homogeneous experimental units cannot be achieved when there is at least a moderately large number of crosses. (The diallel cross is a cross-classified mating design in which a specified number of inbred lines that serve as male parents are crossed with the same lines that serve as female parents. Partial diallel cross experiments are covered in Section 12.9 of John, 1971.)

When incomplete block designs are used, they should ideally be balanced. For example, if there are six treatments but only four could be used in each block and there are six blocks, with the blocks of equal size, we would want each treatment to appear the same number of times in the six blocks combined, and pairs of treatments to appear in blocks the same number of times. If these requirements are met, the design is a *balanced incomplete block* (BIB) *design.*

3.2.1 Balanced Incomplete Block Designs

The model for a BIB design is essentially the same as the model for an RCB design. The model must differ slightly since not all block–treatment combinations are used in the design. Therefore although the model is

$$Y_{ij} = \mu + A_i + B_j + \epsilon_{ij} \qquad i = 1, 2, \ldots, k \quad j = 1, 2, \ldots, t \tag{3.2}$$

which is the same as Eq. (3.1) with the components of the equation thus defined the same way, the model does not apply to all (i, j) combinations, since only a subset of them are used in the experiment. Therefore, the model applies only for the (i, j) combinations that were used.

As with an RCB design, there is the assumption that there is no treatment–block interaction. (We will later see the difficulty in trying to test this assumption, however.)

To illustrate the construction of a BIB design, assume that there are four treatments to be run in blocks but only two treatments can be used in each block. How many blocks are needed for the design to be a BIB design? Since $\binom{4}{2} = 6$, the answer is 6 or a multiple of 6 if more than six blocks could be used. With six blocks, each pair of treatments would occur in a block once (i.e., the design is *balanced* regarding pairs of treatments), as is shown below with A, B, C, and D denoting the treatments.

	Blocks				
1	2	3	4	5	6
A	C	A	A	B	B
B	D	C	D	C	D

This is easy to see but the construction problem is more difficult when there are more than two treatments per block, as then the "pairs" cannot be visualized quite so easily.

Furthermore, it is not always possible to construct a BIB design for a given number of treatments and a given block size. For example, consider Table 3.1 and assume that a block size of 6 is not possible, with 5 being the largest possible block size. Obviously 5 won't work, however, because we won't have six treatments occurring an equal number of times with 20 observations, since it is not a multiple of 6. We would need a minimum of six blocks, not four, for this balance requirement to be met, but the requirement that pairs of treatments occur the same number of times over the blocks would not be met. It would be necessary to drop down to a block size of 4 and use 15 blocks in order for both balance requirements to be met, with each treatment occurring 10 times and every pair of treatments occurring 6 times. This is undoubtedly not intuitively apparent, however, nor should it be.

As with the other designs that are presented in this chapter and with experimental designs in general, we should be mindful of the number of observations that each treatment level mean is computed from, as well as the number of degrees of freedom for the error term. These considerations will generally mandate the use of replicates of the designs in this chapter. We should also be mindful that BIB designs cannot always be constructed for a given number of treatments, replicates of treatments, and block size, as is illustrated in Section 3.2.2.

3.2.1.1 Analysis

The analysis of data from BIB designs is slightly more involved than the analysis of data from RCB designs. As a simple example, consider the BIB design given in the preceding section, with response values as indicated.

	Blocks				
1	2	3	4	5	6
A(8)	C(7)	A(10)	A(6)	B(10)	B(5)
B(4)	D(6)	C(9)	D(6)	C(5)	D(7)

One thing that should be immediately apparent is that we cannot easily separate a block effect from a treatment effect since all treatments do not appear in each block. For example, the total for the third block is more than 50% higher than the total for the first block. Is this because extraneous factors affected the two blocks differently, or does it mean that treatment C has a more pronounced effect on the response than does treatment B? This is, and should be, disturbing because designs in general are not blocked unless the block totals would be expected to differ more than slightly.

It seems as though the treatment totals should be adjusted for differences in the block totals, and the block totals should be adjusted for differences in the treatment totals, which would lead to circular reasoning. As with an RCB design, our interest in the block effect is solely to tell us whether or not blocks should have been used in the experiment; our primary concern is the treatment effect.

The adjustment is made as follows. Let $W_i = 2T_i - B_{(i)}$, with "2" representing the block size, T_i the total for the ith treatment, and $B_{(i)}$ the total of all blocks that contain the ith treatment. (The reason for the "2" is that we are summing only half of the observations in the blocks that contain the ith treatment, whereas $B_{(i)}$ is the total sum of the observations.) For example, $W_{\mathrm{A}} = 2(24) - 43 = 5$. Note that this number is the sum of the differences between each A and the other treatment that is in each block with A. The other adjusted totals are -1, -5, and 1, for B, C, and D, respectively. Note that the sum of the adjusted treatment totals is zero.

This can be explained as follows. Note that $\sum T_i = \mathrm{GT}$, the sum of all the observations, so $\sum 2T_i = 2\mathrm{GT}$. The $\sum B_{(i)}$ also equals 2GT because the block size is 2 and the design is balanced, so that every block total is used twice. The adjusted treatment sum of squares is then obtained as $\sum W_i^2/kt\lambda$, with k and t as previously defined and λ denoting the number of times each pair of treatments occurs together in blocks, which in this case is 1. Thus, $\mathrm{SS}_{\text{treatments(adjusted)}} = 52/8 = 6.5$. Notice that the correction factor $(\mathrm{GT})^2/12$ is not used because a correction is being made in the computation of the $\mathrm{SS}_{\text{treatments(adjusted)}}$.

The other, unadjusted, sums of squares are computed in the usual way, producing the ANOVA table given below.

```
General Linear Model: Y versus Treatments, Blocks

Factor       Type   Levels        Values
Treatments   fixed     4     1, 2, 3, 4
Blocks       fixed     6     1, 2, 3, 4, 5, 6

Analysis of Variance for Y, using Adjusted SS for Tests

Source                 DF    SS      MS      F       P
Treatments (adjusted)   3   6.500   2.167   0.38   0.775
Blocks (unadjusted)     5  19.417   3.883   0.685  0.669
Error                   3  17.000   5.667
Total                  11  42.917
```

The analysis shows that neither blocks nor treatments are significant. This is the usual form of the analysis, although it is potentially misleading regarding the block effect since blocks have not been adjusted. The blocks component must be adjusted to obtain the proper test for blocks. This can be accomplished very easily simply by reversing the order in which the model is specified. For example, in MINITAB the treatment sum of squares was adjusted because the treatment term was the first term specified in the model. Since the first term entered is the one that is adjusted, blocks simply have to be entered first. Doing so produces the following output.

```
Analysis of Variance for Y, using Adjusted SS for Tests

Source                    DF     SS      MS      F      P

Blocks (adjusted)          5  20.333  4.067   0.72   0.653
Treatments (unadjusted)    3   5.583  n.a     n.a    n.a
Error                      3  17.000  17.000  5.667
Total                     11  42.917
```

It can be observed that blocks and treatments are adjusted by the same amount, and in fact it can be shown that

$$\mathrm{SS}_{\text{blocks(adjusted)}} - \mathrm{SS}_{\text{blocks(unadjusted)}} = \mathrm{SS}_{\text{treatments(adjusted)}} - \mathrm{SS}_{\text{treatments(unadjusted)}}$$

This relationship would be an aid in hand computation, although that of course is not recommended other than to become acquainted with what underlies the numbers obtained in the ANOVA table.

3.2.1.2 Recovery of Interblock Information

Blocks are usually considered to be random, although this will not always be the case since in many experiments the blocks that are formed are the only ones that can be formed. When this is the case, the analysis given in the preceding section, which has been termed the *intrablock analysis*, is appropriate. When blocks are random, however, the experimenter has the option of performing an *interblock analysis.* A complete explanation of the latter would be involved and lengthy, and so will not be given here. Perhaps the simplest and most lucid explanation of an interblock analysis is given by Montgomery (2005), which includes a comparison of the treatment effect estimates for both the interblock and intrablock analysis for a particular example. (See also Cochran and Cox (1957, p. 382) and Johnson and Leone (1977).

3.2.1.3 ANOM

Since a BIB design is, as the name indicates, an incomplete design, the correlation structure of the estimators of the treatment means must be determined to see if ANOM can be applied to data obtained from the use of such designs. Nelson (1993) showed that the set of estimators of the treatment means does have the requisite correlation structure. Those estimators do not have the form of the estimators of the treatment means in a CRD or RCB design, however, so Eq. (2.5) cannot be used to obtain the decision lines.

As the reader might surmise, this is due to the fact that the treatment totals are adjusted, as was illustrated in Section 3.2.1.1. Nelson (1993) gave the decision lines for a BIB design as

$$0 \pm h_{\alpha,I,\nu} s \sqrt{\frac{I-1}{IJ(b-1)}} \tag{3.3}$$

with I = the number of treatments, J = the number of blocks, b = block size, and $s = \sqrt{\text{MS}_{\text{error}}}$. Although Nelson (1993) didn't write the expression in this form, when written in this manner we see the close similarity to Eq. (2.5), as the only difference (other than the different symbols) is the $b - 1$ in the denominator of the fraction.

The form for the decision lines given in Eq. (3.3) results from the fact that the form of the estimators of the A_i is given by

$$\widehat{A}_i = \frac{(I-1)}{J(b-1)}\left(w_i - \frac{T_i}{b}\right)$$

where I, J, and b are as previously defined, w_i is the ith treatment total, and T_i is the sum of the block totals in which the ith treatment appears.

As with the RCB design, however, there is no software that will directly produce a BIB ANOM display. Therefore, any such display would have to essentially be produced manually, using Eq. (3.3) to compute the decision limits and computing the average for each treatment (i.e., level of the factor) and displaying the averages on the graph analogous to Figures 2.2 and 2.3.

The use of ANOM for BIB designs is discussed in more detail, with examples, in Nelson et al. (2005, Section 6.3).

3.2.2 Partially Balanced Incomplete Block Designs

Partially balanced incomplete block (PBIB) designs are an obvious, less restrictive, alternative to BIB designs. They are also almost a necessary alternative when the latter cannot be constructed for a given combination of treatments and blocks. (Another, less known alternative is to seek an A- or D-optimal design, using methods such as those given in Reck and Morgan, 2005.) For example, assume that there are six treatments that are to be repeated four times in blocks of size 4, which would obviously require six blocks. To be balanced, each pair of treatments occurs in a block the same number of times, λ, with the latter defined as $r(k-1)/(t-1)$, with r denoting the number of repeats of each treatment, k the block size, and t the number of treatments. Of course, λ must be an integer but here $\lambda = 4(3)/5$, which is not an integer. Clearly, either r or $k - 1$ must be a multiple of $t - 1$ for this example. Although a block size of 5 might be feasible, it might not be possible to use a block size of 10. Similarly, $r = 5$ would not be possible with a block size of 4 for any number of blocks.

A PBIB design has at least two values of λ, such as λ_1 and λ_2, with some pairs of treatments occurring λ_1 times and the other pairs occurring λ_2 times.

Although these designs are less restrictive than BIB designs, they are also less efficient. Specifically, for a BIB design the variance of the difference of two treatment effect estimates will be the same for all pairs of treatments. For a PBIB design with two values of λ, there will be two variances and the average of the two variances is higher than the variance for the BIB design, so the latter is a more efficient design.

PBIB designs are covered in considerable detail in Hinkelmann and Kempthorne (2005), and Chapter 12 of John (1971) is devoted to the subject. The reader is referred to these sources for detailed information on the construction of these designs. Because

BIB designs and PBIB designs cannot be constructed with most statistical software, the use of catalogs of these designs is highly desirable. One such catalog was given by Raghavarao (1971) and an extensive listing of PBIB designs with two associate classes was given by Bose, Clatworthy, and Shrikhande (1954). See also Sinha (1989). Tables of BIB designs were given by Cochran and Cox (1957) and more recently by Colbourn and Dinitz (1996). Nineteen BIB designs were given in Natrella (1963), but unfortunately that has long been out of print.

A recent and most extensive list of BIB designs is given in Hinkelmann and Kempthorne (2005), which lists all known BIB designs and also lists many PBIB designs.

3.2.2.1 Lattice Design

Lattice designs were introduced by Yates (1936) for use in large-scale agricultural experiments and were originally known as quasi-factorial designs. A lattice design, which is a special type of incomplete block design, is related to a Latin square design (covered in Section 3.3) in the sense that two orthogonal Latin square designs (defined in Section 3.4) constitute a simple lattice design. There are different types of lattice designs, but a balanced lattice draws its name from the fact that the treatment numbers can be written at the intersections of lines that form a square lattice. Using results from Kempthorne and Federer (1948), John (1971, p. 262) showed that a lattice design is more efficient than an RCB design that uses the same experimental material. The extent of the efficiency depends, however, on the values of certain variance components, which of course are unknown.

There are various types of lattice designs, including balanced lattices, partially balanced lattices, rectangular lattices, and cubic lattices. A balanced lattice is restrictive in that the number of treatments must be an exact square.

An actual "home improvement" example of a lattice design is given later in Section 3.3.5. Lattice designs are discussed in detail in Chapter 18 of Hinkelmann and Kempthorne (2005), Chapter 10 of Cochran and Cox (1957), and in Federer (1955).

3.2.3 Nonparametric Analysis for Incomplete Block Designs

The analysis of data from most experimental designs is problematic because there generally won't be enough data available to test the assumptions. Consequently, as with statistical procedures in general, it is desirable to perform a nonparametric analysis whenever possible and compare the results with the parametric analysis. Skillings and Mack (1981) gave a nonparametric analysis procedure for data from general incomplete block designs, which was discussed in detail by, for example, Giesbrecht and Gumpertz (2004), and the reader is referred to these sources for information on the approach.

3.2.4 Other Incomplete Block Designs

A class of incomplete block designs known as α-designs was introduced by Patterson and Williams (1976) and found some favor with experimenters. These designs can

be constructed with the Gendex software (see, e.g., http://www.designcomputing.net/gendex).

3.3 LATIN SQUARE DESIGN

Often there will be a need to protect against the possible effect of more than one extraneous factor. For example, in an agricultural experiment there may be soil variation both north/south and east/west relative to a plot of land. There may also be multiple extraneous factors in industrial applications. As discussed in Section 3.1.2, blocking will often be necessary because of processes being out of control. If it is necessary to block on two process characteristics, then a Latin square design could be used. In general, such a design is used when there is a single factor and there are two suspected extraneous factors, the effect of which must be isolated and separated from the error term.

A Latin square, introduced by the famous mathematician Euler in 1783, is as the name implies, a square, an example of which is the following.

```
A  B  C
B  C  A
C  A  B
```

The rows would represent one extraneous factor and the columns the other one, with this 3×3 Latin square thus having three levels of each of the suspected extraneous factors and three levels of the factor of interest, A, B, and C. Notice that each letter occurs once in each column and once in each row, as the design would be unbalanced if this were not the case.

There are 12 ways to construct a 3×3 Latin square such that this requirement is met, as the reader is asked to show in Exercise 3.22. There is only one standard Latin square of this size, however, with a standard Latin square being one that has the letters in the first row and first column in alphabetical order. This should be apparent upon careful inspection of the Latin square given above.

Many of the designs given in this book have been invented in recent years, and virtually all of the designs have been invented within the past 100 years. The Latin square design is an exception, however. As explained by Freeman (2005), the first use of the design (under a different name) may have occurred in 1788, as a French agriculturalist named Cretté de Palluel reported on an experiment that involved the comparison of different types of food for sheep. Sixteen sheep, four from each of four breeds, were fed such that each sheep from each breed received one of four different types of food. The sheep were then weighed at four different times during the winter. Thus, the blocking variables were breed and time. We generally don't think of time as a blocking variable, however, unless the times are close together such as morning and afternoon or consecutive days. So, saying that this is a Latin square design is stretching things a bit, as one could contend that this is a design with a single blocking variable

and multiple response variables corresponding to the four measurement times during the winter.

Since the designation of the levels of the factor (i.e., the letters) is arbitrary, it has been claimed (e.g., Giesbrecht and Gumpertz, 2004, p. 120) that it is necessary to put Latin squares in standard form to see if they really differ, which is why Latin squares are generally written in standard form, at least before any permuting of rows or columns occurs, which is recommended to ensure randomization, as this is the only type of randomization that can be performed. For example, the design

```
A  C  B
B  A  C
C  B  A
```

becomes the same as the previous design if the second and third columns are switched.

It shouldn't make any difference which of these Latin square designs is used if the assumption of no interaction between rows and columns is met. (This assumption is discussed in detail later in the next section.) That assumption won't be met exactly in many applications, however, and then it will make a difference which design is used. Furthermore, the usual recommendation is to take a selected Latin square design and randomly permute rows and columns, then use the design that results from the rearrangements. The method of O'Carroll (1963) might also be used.

Since the assumption of no interaction probably won't be met exactly, it seems highly questionable to say that Latin squares are equivalent if the rows or columns of one design can be permuted so as to create the other design.

The number of Latin squares that are clearly different when in standard form grows rapidly with the order of the square if we think of the order increasing. For example, there are four such squares of order 4, 56 of order 5, and so on.

One obvious restriction of Latin squares, which could be a hindrance, is that the number of levels of all three factors must be the same in order to have a square.

Although selecting a Latin square at random might seem to be a reasonable strategy, this isn't necessarily true. Copeland and Nelson (2000) showed that for Latin squares whose order is 2^k, there is *one* standard square that corresponds to a 2^{3k-k} fractional factorial design, with "2" denoting the levels of each factor, $3k$ denoting the number of factors, and the second k denoting the degree of fractionization. That is, a 2^{3k-k} design is a $1/2^k$ fraction of a 2^{3k} design. (Fractional factorial designs are covered in Chapter 5, as is the relationship between these designs and certain Latin square designs.) From this it follows that not all Latin squares of a given order have equivalent properties.

3.3.1 Assumptions

Another restriction is the assumption that there is no interaction between rows and columns. How often this assumption is violated in practice is conjectural since the true state of nature is never known, but Box, Hunter, and Hunter (2005, p. 160) state

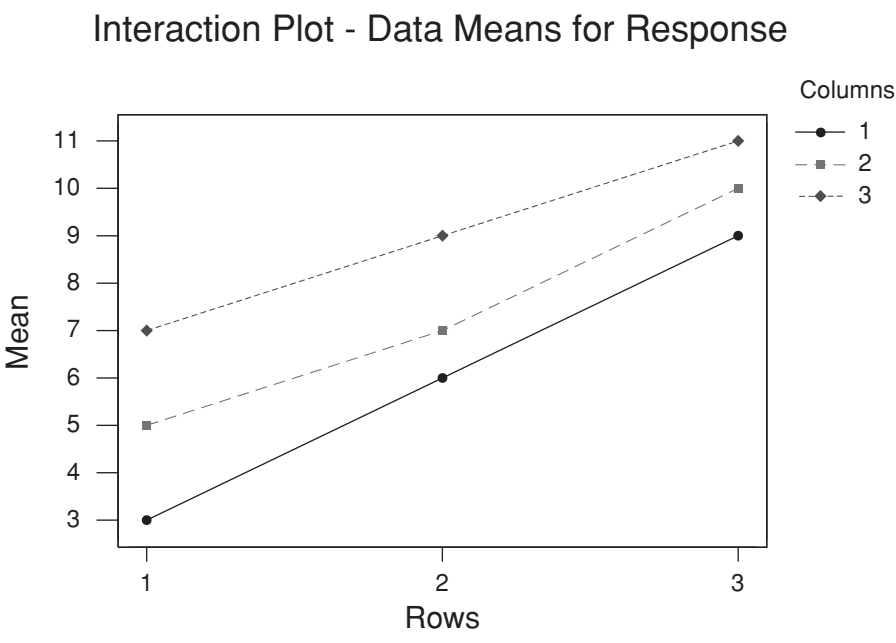

Figure 3.2 A desirable relationship between rows and columns for a Latin square design.

that a Latin square design "has frequently been used inappropriately to study process factors that can interact." Similarly, Nelson et al. (2005, p. 134) stated "... using a Latin square design if the assumption of no interactions were reasonable. Unfortunately, in most physical science and engineering experiments that is not the case, and Latin square designs should be used with great caution since the presence of interactions can invalidate the results."

The assumption means, for example, that if a plant experiment were run on different machines and using different operators, and using these as the row and column blocking factors in an experiment to compare processing methods, that there is no interaction between operators and machines such that, say, one operator seems to work much better on one particular machine than on the other machines. Similarly, for an agricultural experiment the assumption is violated if a noticeable difference in soil fertility occurs in the upper right corner of the design area, or if there is some other irregularity such that a scatterplot of soil fertility, using rows and columns, deviates greatly from parallel lines when the points are connected, as in the graph given in Figure 3.2, for which we might regard soil fertility as the "Response" in that graph.

Another assumption is that there is no interaction between treatments and either rows or columns. This means that if we replaced Rows or Columns in Figure 3.2 with Treatments, we would also have lines that are essentially parallel. (Practically

speaking, parallel lines, which would correspond to zero interaction, would not be likely to occur, just as it is unlikely that a parameter estimate could be equal to the corresponding parameter value.) Jaech (1969) illustrates the consequences of interaction in a Latin square design, which is seen by using expected mean squares. (The latter are discussed in detail in Appendix D to Chapter 4.) The effect of interaction between rows and columns is that the error term is of course inflated, with the consequence that a significant treatment effect may not be detected. See Wilk and Kempthorne (1957) for a detailed discussion of the effect of interactions in a Latin square design.

3.3.2 Model

The model for an $n \times n$ Latin square design is similar to that of a randomized block design, the only difference being that there is an additional blocking variable, so there is an additional term in the model, which is

$$Y_{ijk} = \mu + A_i + R_j + C_k + \epsilon_{ijk} \qquad i = 1, 2, \ldots, t \quad j = 1, 2, \ldots, t \quad k = 1, 2, \ldots, t \tag{3.4}$$

with R_j denoting the effect of the jth row, C_k representing the effect of the kth column, and as for a randomized block design, A_i denoting the effect of the ith treatment. The three subscripts on Y represent the fact that each observation is classified by treatment, row, and column.

3.3.3 Example

A classic example, one that has been used by many writers, is the design of an experiment to compare brands of tires in a driving test, with four cars, four drivers, and four tires of each of four brands available for the experiment. How should the experiment be designed? Obviously all four tires of one brand should not go on each car as then the differences between the brands could not be separated from the car differences. Similarly, each brand should not be solely assigned to wheel position as then brand differences would be confounded with differences due to wheel position.

For the moment we will assume that there is no interest in trying to separate the car effect from the driver effect, then we will address this issue in Section 3.4 in the context of Graeco–Latin squares. By not trying to separate the two (possible) effects, we will actually be measuring the sum of their effects, similar to what happens when fractional factorial designs are used, which are presented in Chapter 5.

With two blocking variables we can use a Latin square design, or perhaps use sets of Latin squares, as is discussed in Section 3.3.5

Assume for the sake of illustration that a single Latin square is to be used, with the following being the one that is randomly chosen (Table 3.2). The numbers in parentheses denote a coded measure of tire wear.

TABLE 3.2 A 4 × 4 Latin Square

	Cars			
Wheel Position	1	2	3	4
LF	A(13)	B(6)	C(12)	D(19)
RF	B(9)	A(9)	D(11)	C(19)
LR	C(9)	D(11)	A(9)	B(8)
RR	D(7)	C(9)	B(3)	A(12)

The analysis of the data is as follows.

```
    Analysis of Variance for Tread Wear

Source          DF       SS       MS      F      P
Brands           3     85.250   28.417   7.41  0.019
Cars             3     92.250   30.750   8.02  0.016
Wheel Position   3     61.250   20.417   5.33  0.040
Error            6     23.000    3.833
Total           15    261.750

S = 1.95789     R-Sq = 91.21%   R-Sq(adj) = 78.03%


Means for Tread Wear

Brands    Mean    SE Mean
1        10.750    0.9789
2         6.500    0.9789
3        12.250    0.9789
4        12.000    0.9789
```

Each of the first three sums of squares in the ANOVA table is computed analogous to the way that the treatment sum of squares is computed for a CRD. That is, for this example SS_{brands} is computed as $\sum_{i=1}^{4} B_i^2/4 - (\sum_{i=1}^{4} B_i)^2/n$, with B_i denoting the total of the observations for brand i, again invoking the general rule that whenever a number is squared in a computing formula, the squared number is divided by the number of observations that add to the number that is squared. As with a CRD, it can be shown that this expression is equivalent to $4\sum_{i=1}^{4}(\bar{B}_i - \bar{B}_{all})^2$, with $\bar{B}_i$ denoting the average of the observations for the ith brand, and $\bar{B}_{all}$ denoting the average of those averages. Since each average is computed from the same number of observations, this is the same as the average of all of the observations.

The sums of squares for Cars and Wheel Position are computed in the same general way, and as stated in Section 2.1.3.3, the computation of the total sum of squares is not design dependent.

The degrees of freedom is $n - 1$ for each of the three factors, and $(n - 1)(n - 2)$ for the error term, with "total" of course having $n^2 - 1$ degrees of freedom.

The analysis shows that 91 percent of the variability in tread wear is accounted for by the model with brands and the two blocking variables. The relatively small p-value for brands shows that there is apparently a difference between brands. Note, however, that the mean for the second brand is far less than the means for the other brands, and there is also a moderate difference between the first and third brand. Yet, the p-value of .019 is not extremely small, which is due to the fact that each average is based on only four observations. (The standard error for each mean of 0.9789 that is given in the output is $\sqrt{\text{MS}_{\text{error}}/4} = \sqrt{3.833/4} = \sqrt{0.958.}$)

The analysis also shows that it was important to use both blocking variables as each is significant at a .05 level of significance (i.e., each p-value is less than .05). Obviously the wheel position is "fixed," as there are four wheel positions on a car, but we would logically regard the cars as a random factor. For a Latin square design the manner of analysis does not depend on whether a blocking factor is fixed or random, although this does matter when various other types of designs are used.

Multiple comparison tests can be used on data from a Latin square design—and the user has the same general options as with the designs given previously—but in this case that is clearly unnecessary as the last part of the output shows that the average tread wear for brand "B" is much less than the average tread wear for the other three brands, which differ very little, so it is obvious why there was a significant result from the F-test.

If, for example, an experimenter wanted to use ANOM, Eq. (2.4) would again be appropriate; it is simply a matter of identifying and specifying the number of observations from which each average is computed, in addition to using the appropriate constant as determined by the number of means that are being compared, the degrees of freedom for the error term, and the significance level. Specifically, for $\alpha = .05$

$$\begin{aligned}\bar{\bar{x}} \pm h_{\alpha,k,\nu}\, s\sqrt{(k-1)/(kn)} &= 10.375 \pm 3.31\sqrt{3.833}\sqrt{3/(4)(4)} \\ &= 10.375 \pm 2.81 \\ &= (7.565,\ 13.185)\end{aligned}$$

The average for the second brand is below 7.565, so we conclude that the mean for that brand is different from the average of all four of the means and the second brand is thus preferable.

Assume that the information on blocking was lost after the experiment was conducted so that only the tread wear figures for each brand were available. Preferably the experiment should be repeated, but assume that a data analyst not knowing that the data were from a Latin square design analyzes the data as having come from a CRD. We can see what the conclusion would have been by adding the sums of squares and degrees of freedom for the two blocking factors to the error degrees of freedom and error sum of squares. It is obvious that doing so would produce a value for the F-statistic of approximately 2, which would result in a p-value well in excess of .05. Thus, a wrong conclusion would have been drawn.

In addition to industrial applications, Latin square designs have been used extensively in medical applications, as indicated by Armitage and Berry (1994). Therefore,

despite the somewhat gloomy picture painted at the end of Section 3.3.1 regarding the likelihood of Latin square assumptions being met, the design has been used successfully in various applications.

3.3.4 Efficiency of a Latin Square Design

Whenever blocking is used with any design, the blocking factor(s) should have a significant effect. Otherwise, the degrees of freedom for the error term is unnecessarily reduced, resulting in a more variable estimator of σ_ϵ^2 than should be the case and reducing the capability of the design to detect differences that are deemed significant. If neither blocking factor in a Latin square design is significant, then the Latin square design is a very inefficient design.

No general conclusion can be drawn about the efficiency of a Latin square design relative to any other design because the efficiency is obviously application- and data dependent. Cochran and Cox (1957, p. 127) gave an expression that would be used for estimating what the error mean square would have been if the blocking variable represented by the rows had not been used, so that the design would have been an RCB design. A similar expression could be derived if the factor represented by the columns had not been used.

3.3.5 Using Multiple Latin Squares

One potentially serious problem when a single Latin square is used is that the number of observations made at each factor level is equal to the number of levels. We should expect this to almost always be an inadequate number to detect a difference between levels of a factor that would be deemed significant, especially when there is only a small number of levels.

This also presents a problem relative to the normality assumption, as that assumption cannot be practically tested when each level mean is computed from only three or four observations.

To illustrate the power problem for the tread wear example given in the preceding section, assume that the experimenter wants to be able to detect a difference in mean tread wear of at least 2 units. With $\hat{\sigma} = 1.96$ in that example, Lenth's calculator (see Section 1.4.4) shows that the probability of detecting a difference of that size is only .23, with the familywise error rate for a set of t-tests controlled at .05.

This problem can be remedied by using multiple Latin squares. Consider the example in Section 3.3.3, with each average that was displayed in the output computed from only four observations since it was a 4×4 Latin square. If two other Latin squares of that size had been randomly selected and combined with the first one, with different rows and columns, the averages would have been computed from 12 observations, which would be much better. Furthermore, the error degrees of freedom is increased to 18, the number of degrees of freedom for the single Latin square times the number of squares.

This increases the power to .68; using four Latin squares of this size produces a power of .81 and five Latin squares gives a power of .88. An experimenter should

use some computing device to obtain these numbers and determine how many Latin squares to use after looking at the power numbers.

The way that degrees of freedom are allocated to the components of the ANOVA table depends upon whether or not there is any relationship between the squares when multiple squares are used.

Example 3.1

To illustrate this, Jaech (1969) gave a nuclear production reactor example involving process tubes, with 10 tubes used and 20 positions on each tube used for test purposes. The 200 tube–position combinations were to be utilized in eight 5×5 Latin squares. The first five tubes were used in square #1 and the second five tubes were used in square #6, but they utilized the same five tube positions, with five different tube positions used for each of squares 2 and 7, 3 and 8, and so on. With this in mind, Jaech (1969) gave the degrees of freedom breakdown for this experiment as follows.

Source	df
Squares	7
Treatments	4
Positions	16
Tubes	8
Error	164
Total	199

The df for squares and for treatments should be obvious since there were eight squares and five treatments, respectively, but the other df may not be obvious. The df for positions is obtained as four degrees of freedom for each of positions 1–5, 6–10, 11–15, and 16–20. The df for tubes consists of four df for each of tubes 1–5 and 6–10. These df should be intuitively apparent when we remember that each Latin square is 5×5. The total df must of course be $200 - 1 = 199$, regardless of how the squares are constructed.

If the squares were unrelated, with randomization performed for each square, there would be no degrees of freedom for positions and tubes. Instead, there would be, as we would expect, 32 df for rows and 32 df for columns, this resulting from 4 df for rows and columns from each of the eight squares. This results in 124 df for the error term—much less than the 164 df under the assumption that the squares are related.

A much smaller df for error will result if the treatment $\times$ square interaction is isolated. Although interactions are assumed negligible when a single Latin square is used, the treatment $\times$ square interaction could be isolated. For this example it would have $(8 - 1)(5 - 1) = 28$ df, so the error would then have 96 df. In general, it is best not to assume that any particular interaction is negligible unless there is a strong reason for doing so. Therefore, it would be desirable to plot the treatment averages against the squares, roughly analogous to Figure 3.1 for an RCB design. That is, for this example there would be $5 \times 8 = 40$ plotted points and the objective would be to

see if the lines that connect averages for each treatment over the squares cross, and if so, how severe is the crossing, as extreme crossing would suggest an interaction that is sufficiently large that it should not be combined with the error term.

The interaction could be tested with a hypothesis test, using the appropriate numbers in the ANOVA table. In general, we would not want to see a significant treatment $\times$ square interaction as this would mean that the values for the factor of interest differ significantly across the squares, something that should theoretically not happen. If this interaction is significant, that should trigger an investigation. If it is not significant, it might be pooled with the error term, as some authors suggest, but the error term will generally have enough degrees of freedom when multiple Latin squares are used that such pooling shouldn't be necessary.

There are various other possibilities for relationships between squares, and these various cases are discussed in considerable detail by Giesbrecht and Gumpertz (2004, pp. 127–134) and the reader is referred to their discussion for additional details.

Although "Squares" is one component of the ANOVA, we would hope this would not be significant. If it is significant, there might be a lurking variable involved, such as some variable that is affecting the response values over time for all three factors. It is interesting to note that the use of multiple Latin squares, although not generally stressed in textbooks on experimental design, was in use as far back as the 1930s. For example, in the famous spindle study of Tippett (1936), two identical Latin square designs were used.

More recently, two mathematics professors at Lafayette College, Gary Gordon and Liz McMahon, came up with a novel application of two orthogonal 4×4 Latin squares. As they describe at http://ww2.lafayette.edu/~math/Gary/Doors.pdf, when they moved into their new house in 1986, each of their two garage doors was white. Deciding to do something colorful to cheer up the neighborhood, they decided to paint their garage doors—in a novel way. Each door had four sections and each section had four raised rectangular panels. Thus, each door had 16 panels, arranged in a 4×4 grid. They selected four colors—purple, light blue, dark blue, and teal—and painted each door in such a way that the doors were orthogonal Latin squares. From the discussion in Section 3.2.2.1 we can also say that their configuration was a lattice design—a somewhat interesting coincidence since the word "lattice" is often used in describing an aspect of a house and means "a window, door, or gate having a lattice," with the latter being "a framework or structure of crossed wood or metal strips." The "rest of the story" can be gleaned at the URL given at the beginning of this discussion.

3.3.6 ANOM

Just as data from an RCB design can be analyzed with ANOM by using decision lines given by Eq. (2.4), so too can data from a Latin square design. As with these other designs, $s = \sqrt{\text{MS}_{\text{error}}}$. Unfortunately, as with data from an RCB design, there is at present no software that can be used to directly generate an ANOM display with data from a Latin square design.

3.4 GRAECO–LATIN SQUARE DESIGN

A variation of a Latin square design that allows for three extraneous factors to be blocked and isolated is called a *Graeco–Latin square design.* Considering the tire experiment in Section 3.3.3, an obvious question to ask is whether or not the experiment should have been designed so that the effect of cars could be separated from the effect of drivers. The answer is "no" because such separation is not of interest. The blocking factor actually measured the sum of those two effects and that was good enough; there would be a problem if an extraneous factor that was thought to be possibly important was not incorporated into the design.

The design derives its name from the fact that Greek letters are typically used to represent the third blocking factor. For example, the first Latin square design given in Section 3.3 can be converted to a Graeco–Latin square design by superimposing a "Graeco square" on top of the Latin square in such a way that each pair of letters occurs only once. The following design is one example.

Aα	Bγ	Cβ
Bβ	Cα	Aγ
Cγ	Aβ	Bα

A more formal way to state how a Graeco–Latin square design is constructed is to say that it is constructed using mutually orthogonal Latin squares of the indicated size, with mutually orthogonal Latin squares being those that produce a design such that each pair of Latin and Greek letters occurs only once, as in the design above. Thus, in order for a Graeco–Latin square design of a given size to exist, a pair of mutually orthogonal Latin squares of that size must exist. No such pair of 6×6 Latin squares exists, so it is not possible to construct a Graeco–Latin square design of that size.

Although it isn't discussed to any extent in the literature, it is possible to view and use a $k \times k$ Graeco–Latin square design as a k^2 design with two blocking factors, with of course $k > 3$. In general, especially with hyper-Graeco–Latin squares, a treatment factor can be substituted for a blocking factor, with the result that the design is a factorial or fractional factorial design with blocking. Factorial and fractional factorial designs are discussed in Chapters 4 and 5, respectively.

3.4.1 Model

The model for the Graeco–Latin square design is just a slight variant of the model for a Latin square design, with the model for the former given by

$$Y_{ijkl} = \mu + A_i + R_j + C_k + G_l + \epsilon_{ijkl} \qquad i = 1, 2, \ldots, t \qquad (3.5)$$
$$j = 1, 2, \ldots, t \quad k = 1, 2, \ldots, t \quad l = 1, 2, \ldots, t$$

with G_l denoting the effect of (the blocking factor level corresponding to) the lth Greek letter and the other components of Eq. (3.5) defined as in Eq. (3.4).

TABLE 3.3 Breakdown of Degrees of Freedom for $t \times t$ Graeco–Latin Square Design

Source	Degrees of Freedom
Row	$t - 1$
Column	$t - 1$
Other Blocking Factor (Greek letters)	$t - 1$
Treatments	$t - 1$
Error (Residual)	$(t - 3)(t - 1)$
Total	$t^2 - 1$

3.4.2 Degrees of Freedom Limitations on the Design Construction

It is easy to construct a 3×3 Graeco–Latin square, and such designs are given in books on design (e.g., Wu and Hamada (2000, p. 74), Kempthorne (1973, p. 187), and Box et al. (2005, p. 161)), with Euler (1782) being the first person to construct the design. Data from experiments in which such designs are used cannot be analyzed using ANOVA—the standard and essentially the only method of a formal analysis in the absence of a prior estimate of σ. This is because there are no degrees of freedom for error, something that must be kept in mind when these types of designs are used. (Of course we might notice that the average response at one or more levels is clearly superior to the average response at other levels, but some type of formal or semiformal testing would clearly be preferable to eyeing the numbers.)

We can see the problem if we look at the degrees of freedom for a $t \times t$ Graeco–Latin square design (Table 3.3).

We see that we "run out" of degrees of freedom by the time we get to the error term when $t = 3$, as all eight degrees of freedom are used by the three blocking factors plus the factor of interest. Thus, unless the design is replicated, the smallest Graeco–Latin square design that permits analysis using ANOVA is a 4×4 design.

Such a design has very little utility when only one square is used, just as is the case with a single Latin square.

(Readers who are returning to this chapter after having read the next chapter might wonder why we can't use methods illustrated in that chapter when there are no degrees of freedom for error. Such methods were developed for unreplicated factorials and it would not be possible, or at least practical, to try to develop a similar method for Latin-square-type designs as the successful use of methods discussed in the next chapter depends upon many effects not being significant.)

The analysis of a Graeco–Latin square is the same as that of a Latin square except that the ANOVA table contains one more blocking factor, represented by the Greek letters, with the sum of squares for it computed in the same general way as the sum of squares for the single factor of interest using the Latin letters.

It is possible to use more than three blocking variables and construct a *hyper-Graeco–Latin* square but such designs will not be discussed here, other than to point out that the degrees of freedom for the error term must also be considered for this design, which does impose a restriction on the size of the design. For example, unless

the design is replicated, a 5×5 design will be necessary if there are four blocking factors, a 6×6 design if there are five blocking factors, and so on. It is easy to show that the number of blocking factors must be one less than the order of the design. Specifically, with s denoting the number of blocking factors, the number of degrees of freedom for the blocking factors will be $s(t-1)$ and the degrees of freedom for the factor of interest is $t-1$. Clearly we must have $s(t-1)+t-1 < t^2-1$. Thus, $(t-1)(s+1) < t^2-1$ will be satisfied only if $s < t$. Thus, the order of the design must be one greater than the number of blocking factors—a severe restriction if there are only a few levels of the factor of interest that are to be considered.

Another potential problem—and a serious one—is that although it may seem desirable to block on *all* expected extraneous sources of variation, the more blocking factors that are used, the more likely the assumption of no interactions is to be violated, and the violation(s) may occur in such a way as to seriously undermine the analysis.

3.4.3 Sets of Graeco–Latin Square Designs

In addition to the degrees of freedom problems with Graeco–Latin square designs noted in the preceding section, such designs have the same problem as a Latin square design regarding the power to detect differences in factor level effects because of the small number of observations from which each factor level average is computed unless a large design is used. Therefore the motivation and need to use sets of Graeco–Latin squares is essentially the same as for a Latin square design, and we would expect that the gains that are achieved in power by using multiple Graeco–Latin squares would parallel the gains realized by using multiple Latin squares discussed in Section 3.3.5.

For example, using the same scenario as was used in discussing single and multiple Latin squares, the power is only .176 when a single Graeco–Latin square design is used, but increases to .489 when two squares are used, to .680 when three squares are used, to .807 when four squares are used, and to .888 when five squares are used. Notice that these are essentially the same values that were obtained for the Latin square design as the number of squares was allowed to increase, as would be expected.

3.4.4 Application

An application of a Graeco–Latin square in the semiconductor industry was described by Michelson and Kimmet (1999). The purpose of the experiment was to evaluate the relationship between the size and density of defects introduced in a wafer fab at specific masking levels and the resultant failures of the individual die at circuit probe for electrical overstress (EOS) related tests.

Defects were intentionally created by placing a large number of known defects on the die during specific levels of manufacturing. The known defects could then be detected at circuit probe EOS testing and linked to the specific process levels.

As Michelson and Kimmet (1999) stated, however, "... the Graeco–Latin square was used as a setup tool rather than as an analysis tool." Specifically, the rows and columns of the 4×4 Graeco–Latin square were simply used to form a grid of 16 cells. Sixteen die were used within each of the cells. The response variables were

the yield of each cell (the count of good die divided by 16), and the count of EOS failures in each cell. The factors within the square were the size of the defects (1, 2, 4, and 8 μm) and the density of the defects (1, 3, 9, and 27 defects/cm^2). Obviously these levels are powers of 2 and 3, respectively, which is presumably how the levels were determined. Thus, unlike a traditional Graeco–Latin square design, here the two factors inside the grid were the ones of interest, rather than having one factor serve as a blocking factor. Thus, this was really a 4^2 factorial with two blocking factors, neither of which was significant.

Since the response variable was percent yield, the variable is thus not normally distributed, so ANOVA is not applicable unless the variable is transformed. The variable was transformed with a logit transformation, however, so the response that was analyzed was actually log [yield/(1 − yield)]. A model was fit that had three terms: size, density, and the size × density interaction. (Note that technically an interaction term cannot be fit (and tested) from an ANOVA standpoint since the interaction term requires nine degrees of freedom and interactions cannot be fit anyway with one observation per cell. A regression approach can be used, however, just as such an approach can be applied to data from a nonorthogonal design, and regression was indeed used in analyzing these data rather than an ANOVA approach.)

Additional Examples

An example for which the design configuration was actually a Graeco–Latin square was described by Cochran and Cox (1957, p. 132), for an experiment described in the literature (Dunlop, 1933), although it was stated that the third blocking factor may have been unnecessary. Box et al. (2005, p. 161) gave an example of a Graeco–Latin square design that is at least realistic, if not real. Overall, however, it is probably safe to say that the use of Graeco–Latin squares has been quite limited.

As stated previously, the more blocking variables that are used, the more likely it is that there will be interactions among them or with the treatment factor. An example of this is the second experiment described by Sheesley (1985). There was a single factor of interest and three factors that were viewed as blocking factors. The data were analyzed as having come from a $2^3 \times 3$ design with four replications. It was a good thing that it was not run as a Graeco–Latin square design because there were some moderately large interactions among the blocking factors. This was not a problem because of the way that the data were analyzed, however, and Sheesley (1985) stated, "The significance of the blocking/process variables and interactions among them is not of concern since the purpose of the test was to simply account for their existence in order to evaluate the effect of lead types."

This is undoubtedly a common occurrence and militates against the use of Latin square type designs with more than two blocking factors.

3.4.5 ANOM

As with a Latin square design, data from a Graeco–Latin square design can be analyzed with ANOM by using the decision lines given by Eq. (2.4), but there is apparently

no software that will directly construct the display. From a practical standpoint, there would be a problem even if such software did exist since it is virtually impossible to check for normality of populations with one or a few observations from each population. Specifically, assume that a single 3×3 Graeco–Latin square design is used so that the mean of each level of the factor is computed from three observations. There is no way to assess the normality assumption that underlies ANOM with only three observations. Of course this problem must also be addressed when ANOVA is used, which is one reason why multiple Graeco–Latin squares should generally be used rather than a single Graeco–Latin square.

3.5 YOUDEN SQUARES

In addition to the small number of degrees of freedom for the error term, another major shortcoming for nonagricultural use is the fact that the number of treatments and levels of the two blocking factors must be the same. (Obviously this is not a problem for agricultural experiments in which soil fertility in the two directions forms the blocking factors and the number of plots of land could be easily selected to equal the square of the number of treatments.)

This restriction will be unrealistic in many industrial experiments, however. Consider, for example, the tread wear experiment of Section 3.3.3. Obviously the number of wheel positions is fixed, but perhaps there are only three tire brands under consideration instead of four. Or perhaps only three cars are available for the experiment instead of four.

A prominent industrial statistician, W. Jack Youden (1900–1971), once was confronted with a situation where a Latin square design would seem to be appropriate because of the need to use two blocking factors, but there was a physical limitation that prevented the use of a square. This led him to use selected rows of a Latin square design. Such a design might be called an incomplete Latin square, but was termed a Youden square by R. A. Fisher. This is discussed in Youden (1937), which is apparently the first presentation of a Youden design. Most of the published Youden squares are due to Youden.

The term "Youden square" is of course a misnomer because if we use part of a Latin square, we no longer have a layout that is square. Therefore, we will henceforth refer to the type of design as the "Youden design."

An obvious question is "Can the rows of a Latin square design be selected at random to produce a Youden design, or can only certain rows of a given Latin square design be used?" The rows cannot be selected at random in arriving at a desired number of rows as there are requirements relating to balance that must be met. This is because a Youden design is essentially a BIB design. That is, the columns become incomplete blocks since not all of the rows are used. In order to have balance, each pair of treatments must occur the same number of times, and of course no treatment occurs more than once in a column, with not all treatments occurring in any column. That is, the columns are balanced incomplete blocks, but a BIB design doesn't exist for all sizes of rectangular configurations.

To use a simple example as an illustration, if we start with a 3×3 Latin square and delete the last row, we have six observations and the design is as follows.

A	B	C
B	C	A

The columns will form incomplete blocks with two treatments per block. In order for the balance requirement to be met, the number of pairs of treatments, which is obviously $\binom{3}{2} = 3$, must occur the same number of times per block over the set of blocks. Notice that each of the pairs (AB, AC, and BC) occurs once, so the balance requirement is met.

Similarly, if we start with the 4×4 Latin square in Table 3.2 and delete the last row, we will also obtain a Youden design with each pair of treatments occurring twice. (Notice that this balance requirement is also met if we delete any one row from that design.)

From these simple examples it might *seem* as though we could delete any row from a Latin square design and obtain a Youden design and that is true. The balance property can be easily established when the Latin square is a 4×4 design, and the reader is asked to establish the result in general in Exercise 3.5. For the 4×4 design, since one row is deleted, each block contains three treatments, and since the deleted row contains different treatments, any pair of blocks in the Youden design will have two treatments in common. Hence every pair of treatments will occur in two blocks, thus satisfying the balance property.

This won't necessarily be true for larger designs, however. For example, Dean and Voss (1999) point out that there is no BIB design for eight treatments in eight blocks of size 3. Therefore, it follows that we cannot delete *any* group of five rows from an 8×8 Latin square and obtain a Youden design. Thus, since it isn't always possible to arbitrarily delete multiple rows from a Latin square and produce a Youden design, catalogs of Youden designs are useful, and these are discussed in Section 3.5.2.

3.5.1 Model

Since a Youden design is a BIB design, the model for a Youden design is essentially the same as the model for a BIB design. We will write it slightly differently, however, to conform to the notation for a Latin square design since it is also an incomplete Latin square. The model is

$$\begin{aligned} Y_{ijk} &= \mu + A_i + R_j + C_k + \epsilon_{ijk} \\ & i = 1, 2, \ldots, t \quad j = 1, 2, \ldots, m \quad k = 1, 2, \ldots, t \end{aligned} \tag{3.6}$$

with $m < t$, and Eq. (3.6) is otherwise the same as Eq. (3.4).

3.5.2 Lists of Youden Designs

Since most Youden designs cannot be simply constructed, it will generally be necessary to refer to lists/catalogs of such designs. Unfortunately such lists are not readily available. Natrella (1963) listed 29 such plans and gave the actual design for 9 of them, but unfortunately that book has long been out of print, as noted previously, and its Internet successor, the *NIST/SEMATECH e-Handbook of Statistical Methods* at http://www.itl.nist.gov/div898/handbook, does not cover Youden designs.

Cochran and Cox (1957), which is in print, gives more of these designs than did Natrella (1963), however, and lists the actual design layout for almost all of these 29 plans. At present this is the best source for Youden designs.

3.5.3 Using Replicated Youden Designs

Just as there will generally be a need to use multiple Latin squares, there will also generally be a need to use a replicated Youden design. Unlike the case with Latin squares, however, there will generally not be multiple Youden designs from which to select. Therefore, it will usually be necessary to replicate a given Youden design. The Java applets used previously do not have the capability to perform power calculations for a Youden design. Since a Youden design is really a BIB design, as stated previously, one can use the same approach as is used for determining a suitable sample size for that design. This is discussed by Dean and Voss (1999, p. 407), who illustrate how to solve for the number of replicates.

3.5.4 Analysis

Not surprisingly, the analysis of data from a Youden design is the same as the analysis for a BIB design, which was illustrated in Section 3.2.1.1. Therefore, the reader is referred to that section for details. The analysis using ANOM is of course also the same, and the application of ANOM to Youden squares is described in Section 6.4 of Nelson et al. (2005).

3.6 MISSING VALUES

Missing values create a major problem when a design in this chapter is used. In particular, a missing value in a Latin square causes the "square" to be lost. In general, missing values cause a major problem whenever a design is constructed for more than a single factor. Of course, a value can also become "missing" when there is a botched experimental run and the recorded value is nonsensical.

Assume that an observation is missing from an RCB design. One approach is to estimate the missing value by minimizing the error sum of squares, using a standard calculus approach. The problem with this method is that it biases the results because the value that is substituted is the value that most closely fits the model that is used. The actual value will almost certainly be different and will not adhere as closely to the fitted model as would the imputed value.

An unsatisfactory approach would be to delete the observations on the same treatment that is missing in one block from the other blocks. This would result in fewer treatments being compared and would defeat the purpose of the experiment.

A better approach is to analyze the unbalanced data that results when there are missing observations. An RCB design becomes "incomplete" when there is a single missing observation and also becomes unbalanced because the block sizes will be different. An "adjusted analysis" can be performed, similar to what is done with a BIB design, while recognizing that the power of the design is lessened by the missing observation.

For example, assume that we have the following example, with A, B, C, and D denoting the four treatments and the numbers in parentheses denoting the response values, with "*" denoting a missing value.

Blocks		
1	2	3
C(10)	E(16)	A(19)
D(12)	C(18)	E(18)
E(14)	D(20)	D(11)
B(16)	A(21)	C(10)
A(13)	B(24)	B(*)

A "standard" analysis of these data as a two-way classification of the data would be met with an error message. Instead, the data must be analyzed as unbalanced data, which can be done in MINITAB, for example, by using the general linear model capability. The output for this example is shown below, indicating that the blocking was needed but there is no significant treatment effect. Note that none of the data are discarded as the total df is 13.

```
General Linear Model: Y versus Blocks, Treatments

Factor      Type     Levels     Values
Blocks      random      3       1,2,3
Treatments  fixed       5      1, 2, 3, 4, 5

Analysis of Variance for Response, using Adjusted SS for Tests

Source      DF    Seq SS    Adj SS   Adj MS    F      P
Blocks       2   125.914   118.267   59.133  8.67  0.013
Treatments   4    74.067    74.067   18.517  2.72  0.118
Error        7    47.733    47.733    6.819
Total       13   247.714

      S = 2.61133   R-Sq = 80.73%   R-Sq (adj) = 64.21%
```

As a second example, consider the example in Section 3.1.4 and assume that the last observation (i.e., in the lower right corner) is missing. The analysis is given below. (As indicated, blocks are assumed to be random but that classification has no effect

on the analysis since the error term is really the interaction and the random effect is tested against the error in a two-factor mixed model, with the fixed effect tested against the interaction. But since interaction and error are inseparable, both effects are tested against error.)

```
General Linear Model: Conversion Gain versus Test Set, Resistor

    Factor    Type    Levels    Values
    Test Set  fixed      4     1,2,3,4
    Resistor  random     6     1,2,3,4,5,6

Analysis of Variance for Conversion Gain, using Adjusted SS for Tests

   Source      DF   Seq SS   Adj SS    Adj MS     F       P
   Test Set     3   902.67   893.33    297.78  309.68   0.000
   Resistor     5     5.59     5.59      1.12    1.16   0.375
   Error       14    13.46    13.46      0.96
   Total       22   921.72

   S = 0.980585     R-Sq = 98.54%       R-Sq(adj) = 97.70%
```

Notice that the loss of the observation has only a very slight effect on the numerical results and the conclusions are unaffected.

Similarly, the analysis of data from a Latin square design with a single missing observation is also straightforward. The computer output below is for the example in Section 3.3.3, with the observation in the lower right corner of the square assumed to be missing.

```
General Linear Model: Tread Wear versus Brands, Cars, Wheel Position

Factor           Type    Levels    Values
Brands          fixed       4      1,2,3,4
Cars            random      4      1,2,3,4
Wheel Position  fixed       4      1,2,3,4

Analysis of Variance for Tread Wear, using Adjusted SS for Tests

Source          DF    Seq SS   Adj SS   Adj MS    F      P
Brands           3    84.517   85.389    28.463   6.23  0.038
Cars             3   104.861   65.722    21.907   4.80  0.062
Wheel Position   3    46.722   46.722    15.574   3.41  0.110
Error            5    22.833   22.833     4.567
Total           14   258.933

S = 2.13698   R-Sq = 91.18%       R-Sq(adj) = 75.31%
```

With only 16 original observations, we might expect that removing one observation could make a noticeable difference, and we see that it does happen here. In particular,

there are major changes in the p-values with the p-value for Brands exactly doubled and a greater percentage p-value increase occurring for Cars and Wheel Position. Here it would have been better to simply repeat the experiment, if possible and practical, than to try to analyze the data with the missing value. In general, the percentage of missing observations should be kept in mind before trying to draw conclusions from the remaining observations.

The analysis by ANOVA would proceed similarly for other designs discussed in this chapter, such as Graeco–Latin squares and Youden designs, when there is missing data.

Missing data do create major problems when ANOM is used, however, because even one missing data point destroys the equicorrelation structure that is necessary for ANOM to be used. Subramani (1992) proposed a step-by-step ANOM approach for use with missing data and illustrated the technique by applying it to data obtained from the use of a randomized block design.

3.7 SOFTWARE

Experimental design software is generally not oriented toward Latin square designs and its variants. For example, Latin square designs are not explicitly incorporated into Design-Expert and the user has to do some work to construct one, as is shown in a help file, which is discussed later in this section. Similarly, a Latin square design cannot be constructed using MINITAB, although a help file does explain how the model should be specified in MINITAB so as to allow data from such a design to be analyzed. The design also cannot be constructed directly using JMP, but can be constructed indirectly as a D-optimal design with one k-level factor and two blocks. (Optimal designs are covered in Section 13.7.) A Latin square design can be constructed using SAS, although SAS code must be written to accomplish this. (SAS code for randomly selecting a 5×5 Latin square is given by Giesbrecht and Gumpertz, 2004, p. 121.)

As stated, a Latin square design also cannot be constructed directly using Design-Expert, but can be constructed with some work. That is, as with the other software, it is not possible to select the design from a menu of designs. Rather, it is necessary to build the design using the general factorial design capability and specifying three categorical factors—for rows, columns, and treatments. The steps that must be followed in creating a 5×5 Latin square design and analyzing the resultant data are given at http://www.statease.com/rocket.html. In essence, a 5^3 design is initially constructed, with the last 100 rows of the design then being deleted. It might thus look as though a 5^{3-1} design is being constructed, but that is not the case as the Latin square is actually entered manually. The row and column factors that are specified simply serve to provide the template for the design construction. Since the design (i.e., the letters for the treatments, A–E) is entered manually, it would be a bit of a stretch to state that the software is constructing the design. Furthermore, a Latin square is obviously not randomly selected when the user is instructed to enter a particular Latin square. Although the steps are easy to follow, most software users would undoubtedly prefer to be able to construct the design directly and in fewer steps, and without having to actually enter the design themselves.

There is clearly a need for software developers to devote some attention to Latin square designs and their variants. It is relatively easy to go to a catalog of designs and randomly select a Latin square design of a particular size, but multiple Latin squares will generally be needed, which requires a mechanism for selecting them and software to analyze the data.

The situation is similar, although not as dire, regarding randomized block designs. The latter is essentially a two-factor design with no interaction and with randomization within each level of the blocking factor. Since the design layout is essentially that of a replicated one-factor design, the design can be easily constructed in Design-Expert, although as with a Latin square design, there is no menu that lists "randomized block design" as one of the menu items. Similarly there is no such menu item when JMP is used, but the design can easily be constructed by specifying one of the two factors to be a blocking factor.

The situation is different with incomplete block designs as such designs cannot be constructed indirectly. Specifically, a BIB design will be balanced only if it is constructed so as to be balanced, and similarly for PBIB designs. Balanced incomplete block and partially balanced incomplete block designs cannot be directly constructed with JMP (but can be constructed indirectly), MINITAB, or Design-Expert. Balanced incomplete block designs can be constructed using PROC OPTEX in SAS/QC, however, with appropriate code, including a BLOCKS statement. As stated in http://support.sas.com/techsup/faq/stat_key/k_z.html, "No procedure creates these specifically, but SAS/QC PROC OPTEX may find such designs if they are optimal according to the criterion used."

There *is*, however, statistical software that can be used to construct the designs given in this chapter. For example, Statgraphics can be used to construct RCB designs, BIB designs, Latin squares, Graeco–Latin squares, and hyper-Graeco–Latin squares, and Gendex, which is a DOE toolkit, will create certain types of incomplete block designs, including α-designs.

Although construction of incomplete block designs with most software packages is obviously a problem, there are various statistical packages that will analyze data from such designs. The extent of the use of incomplete block designs today seems questionable, however, at least in certain fields. For example, John (2003) stated "... incomplete block designs—a mathematical subject of theoretical use in agriculture and, so far as I can tell, of little interest to engineers." Incomplete block designs are generally not taught to engineers in short courses, so this may be one of the primary reasons for the fact that such designs are not used to any extent in engineering experimentation.

3.8 SUMMARY

It is important to isolate and extract extraneous factors from the error term, so that the latter is not improperly inflated, with the consequence that the factor(s) of interest may not be judged significant, as was discussed and illustrated in Section 3.3.3.

There are potential problems involved in the use of the designs given in this chapter, however, as no interactions are permitted for these designs. This assumption should always be checked graphically, as was illustrated in Figure 3.1. Statistical tests are also available for testing specific types of interactions (see Hinkelmann and Kempthorne, 1994).

The assumption of normality also presents a problem, especially when a single Latin square or Graeco–Latin square design is used, as it is not practical to test for normality with only a few observations. This is another reason why multiple Latin squares, Graeco–Latin squares, and hyper-Graeco–Latin squares should be used.

It is also important to bear in mind degrees of freedom restrictions and considerations, which will generally lead to the use of multiple Latin squares, Graeco–Latin squares, or hyper-Graeco–Latin squares. For additional reading on Latin squares and their variations, the reader is referred to Chapter 6 of Giesbrecht and Gumpertz (2004); for further reading on incomplete block designs the reader is referred to Chapter 11 of Dean and Voss (1999) and Chapter 8 of Giesbrecht and Gumpertz (2004). See also Montgomery (2005). Also worth noting is the Internet project at www.designtheory.org, which is intended to eventually be a source and catalog of a large number of designs, and the Web site includes software that is available as freeware. Much of what is there and will be there in the future will be of interest only to researchers in design theory, however, such as the 26×26 Latin square design that can be found there.

Finally, the general untestable assumption of normality underlies an ANOVA approach for each of these designs. Consequently, it would be desirable to use a nonparametric approach and compare the results, but other than the nonparametric approach given by Skillings and Mack (1981) for incomplete block designs that was mentioned in Section 3.2.3, nonparametric approaches for designs with blocking factors seem to be almost nonexistent.

REFERENCES

Armitage, P. and G. Berry (1994). *Statistical Methods in Medical Research*, 3rd ed. Oxford, UK: Blackwell.

Bose, R. C., W. H. Clatworthy, and S. S. Shrikhande (1954). Tables of partially balanced designs with two associate classes. Technical Bulletin #107, North Carolina Agricultural Experiment Station.

Box, G. E. P., S. Bisgaard, and C. Fung (1990). *Designing Industrial Experiments.* Madison, WI: BBBF Books.

Box, G. E. P., J. S. Hunter, and W. G. Hunter (2005). *Statistics for Experimenters: Design, Innovation and Discovery.* Hoboken, NJ: Wiley.

Cochran, W. G. and G. M. Cox (1957). *Experimental Designs*, 2nd ed. New York: Wiley.

Colbourn, C. J. and J. H. Dinitz, eds. (1996). *The CRC Handbook of Combinatorial Designs.* Boca Raton, FL: CRC Press.

Copeland, K. A. F. and P. R. Nelson (2000). Latin squares and two-level fractional factorial designs. *Journal of Quality Technology*, **32**(4), 432–439.

Dean, A. and D. Voss (1999). *Design and Analysis of Experiments.* New York: Springer-Verlag.

Dunlop, G. (1933). Methods of experimentation in animal nutrition. *Journal of Agricultural Science*, **23**, 580–614.

Euler, L. (1782). Reserches Sur Une Novelle Espece des Quarres Magiques. Verh. Zeeuwsch. Genooot. *Weten Vliss*, **9**, 85–239.

Federer, W. T. (1955). *Experimental Design: Theory and Application.* New York: MacMillan.

Freeman, G. (2005). Latin squares and su doku. *Significance*, **2**(3), 119–122.

Giesbrecht, F. G. and M. L. Gumpertz (2004). *Planning, Construction, and Statistical Analysis of Comparative Experiments.* Hoboken, NJ: Wiley.

Hinkelmann, K. and O. Kempthorne (1994). *Design and Analysis of Experiments. Vol. 1: Introduction to Experimental Design.* New York: Wiley.

Hinkelmann, K. and O. Kempthorne (2005). *Design and Analysis of Experiments. Vol. 2: Advanced Experimental Design.* Hoboken, NJ: Wiley.

Jaech, J. L. (1969). The Latin square. *Journal of Quality Technology*, **1**(4), 242–255.

John, P. W. M. (1971). *Statistical Design and Analysis of Experiments.* New York: MacMillan (reprinted in 1998 by the Society for Industrial and Applied Mathematics).

John, P. W. M. (2003). Plenary presentation at the *2003 Quality and Productivity Research Conference*, IBM T. J. Watson Research Center, Yorktown Heights, NY, May 21–23, 2003. (The talk is available at http://www.research.ibm.com/stat/qprc/papers/Peter_John.pdf.)

Johnson, N. L. and F. C. Leone (1977). The recovery of interblock information in balanced incomplete block designs. *Journal of Quality Technology*, **9**(4), 182–187.

Kempthorne, O. (1973). *Design and Analysis of Experiments.* New York: Robert E. Krieger Publishing Co. (copyright held by John Wiley & Sons, Inc.)

Kempthorne, O. and W. T. Federer (1948). The general theory of prime power lattice designs. *Biometrics*, **4** (Part I), 54–79; (Part II), 109–121.

Mead, R., R. N. Curnow, and A. M. Hasted (1993). *Statistical Methods in Agriculture and Experimental Biology.* London: Chapman & Hall.

Michelson, D. K. and S. Kimmet (1999). An application of Graeco–Latin square designs in the semiconductor industry. In *Proceedings of the American Statistical Association*, American Statistical Association, Alexandria, VA.

Montgomery, D. C. (2005). *Design and Analysis of Experiments*, 6th ed. Hoboken, NJ: Wiley.

Natrella, M. (1963). *Experimental Statistics*, National Bureau of Standards Handbook 91. Washington, DC: United States Department of Commerce.

Nelson, P. R. (1993). Additional uses for the analysis of means and extended tables of critical values. *Technometrics*, **35**(1), 61–71.

Nelson, P. R., M. Coffin, and K. A. F. Copeland (2003). *Introductory Statistics for Engineering Experimentation.* San Diego, CA: Academic Press.

Nelson, P. R., P. S. Wludyka, and K. A. F. Copeland (2005). *The Analysis of Means: A Graphical Method for Comparing Means, Rates, and Proportions.* Philadelphia: Society for Industrial and Applied Mathematics (co-published with the American Statistical Association, Alexandria, VA).

Nyachoti, C. M., J. D. House, B. A. Slominski, and I. R. Seddon (2005). Energy and nutrient digestibilities in wheat dried distillers' grains with solubles fed to growing pigs. *Journal of the Science of Food and Agriculture*, **85**(15), 2581–2586.

O'Carroll, F. (1963). A method of generating randomized Latin squares. *Biometrics*, **19**, 652–653.

Patterson, H. D. and E. R. Williams (1976). A new class of resolvable incomplete block designs. *Biometrika*, **63**, 83–92.

Peake, R. (1953). Planning an experiment in a cotton spinning mill. *Applied Statistics*, **2**, 184–192.

Raghavarao, D. (1971). *Constructions and Combinatorial Problems in Design of Experiments.* New York: Wiley. (published in paperback in 1988 by Dover Publications).

Reck, B. and J. P. Morgan (2005). Optimal design in irregular BIBD settings. *Journal of Statistical Planning and Inference*, **129**, 59–84.

Sarrazin, P., A. F. Mustafa, P. Y. Chouinard, and S. A. Sotocinal (2004). Performance of dairy cows fed roasted sunflower seed. *Journal of the Science of Food and Agriculture*, **84**(10), 1179–1185.

Sheesley, J. H. (1985). Use of factorial designs in the development of lighting products. In *Experiments in Industry: Design, Analysis, and Interpretation of Results*, pp. 47–57 (R. D. Snee, L. B. Hare, and J. R. Trout, eds.). Milwaukee, WI: Quality Press.

Singh, M. and K. Hinkelmann (1999). Analysis of partial diallel crosses in incomplete blocks. *Biometrical Journal*, **40**(2), 165–181.

Sinha, K. (1989). A method of constructing PBIB designs. *Journal of the Indian Society of Agricultural Statistics*, **41**, 313–315.

Skillings, J. H. and G. A. Mack (1981). On the use of a Friedman-type statistic in balanced and unbalanced block designs. *Technometrics*, **23**(2), 171–177.

Subramani, J. (1992). Analysis of means for experimental designs with missing observations. *Communications in Statistics—Theory and Methods*, **21**(7), 2045–2057.

Tippett, L. H. C. (1936). Applications of statistical methods to the control of quality in industrial production. *Transactions of the Manchester Statistical Society*, Session 1935–1936, 1–32.

van der Linden, W. J., B. P. Veldkamp, and J. E. Carlson (2004). Optimizing balanced incomplete block designs for educational assessment. *Applied Psychological Measurement*, **28**(5), 317–331.

Wilk, M. B. and O. Kempthorne (1957). Non-additivities in a Latin square design. *Journal of the American Statistical Association*, **52**, 218–236.

Wu, C. F. J. and M. Hamada (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization.* New York: Wiley.

Yates, F. (1936). A new method of arranging variety trials involving a large number of varieties. *Journal of Agricultural Science*, **26**, 424–455.

Youden, W. J. (1937). Use of incomplete block replications in estimating tobacco-mosaic virus. *Contributions from Boyce Thompson Institute*, **9**(1), 41–48 (reprinted in the January, 1972 issue of the *Journal of Quality Technology*).

EXERCISES

3.1 Assume that a randomized complete block (RCB) design was used and either (a) the block totals were approximately the same but the treatment totals differed considerably, or (b) the treatment totals were approximately the same but

the block totals differed considerably. Does either (a) or (b) suggest that this type of design should have been used? Explain.

3.2 Assume that you are given a 4×4 hyper-Graeco–Latin square design to analyze with four blocking factors. What would you tell the experimenter who gave you the data?

3.3 Explain how an RCB design for five factors and four blocks should be constructed.

3.4 Assume that an experimenter deletes two rows of a 3×3 Latin square and claims to have constructed a Youden design. Do you agree that this is a Youden design or is there another name for the design? Explain.

3.5 Prove that the properties of a Youden design will be met whenever the design is formed by deleting one row of a Latin square.

3.6 Consider the following experiment, a modification of one found on the Internet. An engineer is studying the effect of five illumination levels on the occurrence of defects in an assembly operation. Because time may be a factor in the experiment, he decided to run the experiment in five blocks, with each block corresponding to a day of the week. The department in which the experiment is conducted has (only) four workstations and these stations represent a potential source of variability. The engineer decided to run an experiment with the layout given below, with the rows representing days, the columns representing workstations, and five treatments denoted by the letters A through D. The data, coded for simplicity, are shown below.

	Work Station			
Day	1	2	3	4
1	A(6)	B(1)	C(2)	D(3)
2	B(2)	C(2)	D(3)	E(7)
3	C(4)	D(5)	E(4)	A(3)
4	D(7)	E(6)	A(4)	B(2)
5	E(5)	A(2)	B(2)	C(5)

(a) What type of design was used?
(b) Perform the appropriate analysis, but first state what assumptions must be made and check those assumptions.
(c) What is your conclusion regarding the five illumination levels?

3.7 The following problem can be found on the Internet: "The following experiment was designed to find out to what extent a particular type of fabric gave homogeneous results over its surface in a standard wear test. In a single run the test machine could accommodate four samples of fabric, at positions 1, 2,

3, and 4. On a large sheet of the fabric, four areas A, B, C, and D were marked out at random at different places over the surface. From each area four samples were taken, and the sixteen samples thus obtained were compared in the machine with the following results, given in milligrams of wear." The design layout was a Latin square with the test runs denoting the columns and the rows denoting the positions. What is the factor of interest and what are the blocking factors?

3.8 **(a)** Construct a 5×5 Latin square design.
(b) Write out the model when this design is used and state the assumptions that must be made, then state how you would check those assumptions.
(c) Describe an application in your field of study in which this design might be successfully employed.

3.9 Assume that an experiment is to be conducted to examine the effect of four different gasoline additives A, B, C, and D on reduction of nitrogen oxides. Four cars are available for the experiment, as are four drivers. Assume that differences are expected between the cars and between the drivers, so these are to serve as blocking factors.

(a) How would you design the experiment if 16 runs are to be made, assuming an identical driving course (terrain)?
(b) Now assume that because of time considerations, it isn't practical to use the same driving course for each run, as that would require various combinations of cars and drivers to be constantly "in waiting," so this would not be an efficient use of resources. To alleviate this problem, four driving courses (routes) are laid out. Now explain how the experiment should be designed and performed.
(c) Considering the number of observations, would you suggest that the design that you named in part (b) be replicated? If so, how would the replication be performed?

3.10 Assume that a 3×3 Latin square design is to be used and a particular square is randomly selected. Before the experiment is conducted, however, the experimenter decides that one of the blocking factors is actually a factor of interest. With this is mind, how should the design be described and should the analysis of the data be affected by this decision? Explain.

3.11 Nelson (1993, references) gave an example to illustrate the application of data from an experiment with a Graeco–Latin square design. A manufacturer of disk drives was interested in studying the effect of four substrates (aluminum, nickel-plated, and two types of glass) on the amplitude of the signal that is received. There were four machines, four operators, and four days of production that were to be involved, with machines, operators, and days to serve as blocking variables. The 4×4 Graeco–Latin square design is given below, with

columns corresponding to operators, rows representing machines, Greek letters representing days, and Latin letters representing the substrates. The numbers in parentheses are the coded response values.

Aα (8)	Cγ (11)	Dδ (2)	Bβ (8)
Cδ (7)	Aβ (5)	Bα (2)	Dγ (4)
Dβ (3)	Bδ (9)	Aγ (7)	Cα (9)
Bγ(4)	Dα (5)	Cβ (9)	Aδ (3)

(a) State the assumptions that must be made if ANOM is to be applied to these data. Do these assumptions appear to be met? Explain.

(b) If the assumptions appear to be met, analyze the data with ANOM by comparing the substrate means against the decision lines (without a display).

3.12 Assume that one or more 4×4 Latin squares are to be used and σ is known to be approximately 1. If the power is to be .95 of detecting a difference between two means of the factor that is at least 2 units using the Scheffé approach, how many Latin squares should be used?

3.13 Construct an example of a 3×3 Latin square with data for which the sum of squares for the factor of interest is zero.

3.14 The abstract of an article in the journal *Arthritis and Rheumatism* (**50**(2), 458–468, February 2004) contains the following sentence: "Forty-two physical signs and techniques were evaluated using a 6×6 Latin square design." Is there anything wrong with that statement? Read the article and determine if the analysis of the data and inference drawn therefrom are correctly performed. If so, explain the analysis relative to the quote given above. If the analysis and/or inference are incorrectly performed, perform the correct analysis, if possible. If a proper analysis is not possible, explain why not.

3.15 Assume that a 5×5 Latin square is used and one of the graphical checks on the assumptions shows a moderate columns $\times$ treatments interaction. Explain how this could affect the determination of whether there is a treatment effect.

3.16 Assume that a Latin square design was used and part of the ANOVA table is as follows.

ANOVA

Source	DF	SS
Rows	3	122.4
Columns	3	26.9
Treatments	3	167.8
Error	6	61.2

Do the numbers suggest that a Latin square design should have been used? Explain. If not, what design could have perhaps been used instead?

3.17 In his article "The Latin square" in the October 1969 issue of the *Journal of Quality Technology*, J. L. Jaech gave the following data, in parentheses, for a Latin square design example, but with data that were not generated using the model for a Latin square.

C(19)	A(9)	B(−4)	D(9)
B(13)	D(9)	A(6)	C(16)
A(18)	C(15)	D(3)	B(8)
D(10)	B(11)	C(11)	A(12)

What model assumption(s) for a Latin square design appears to be violated, if any? Explain.

3.18 If a randomized block design would be useful in your field of engineering or science, give an example. If such a design would not be useful, explain why.

3.19 Assume that there is a need to detect fairly small differences between treatments, and the BIB design given at the start of Section 3.2.1.1 is to be used. What would you suggest to the experimenters?

3.20 **(a)** Complete the following ANOVA table for a Latin square design, assuming that all assumptions are met.

```
 Analysis of Variance for Yield

Source      DF      SS     MS       F

Rows         4      12
Columns             16
Treatments
Error               36
Total              112
```

(b) Use $\alpha = .05$ and test the hypothesis that the treatment means are equal.
(c) Assume for the moment that the hypothesis is not rejected. What would you recommend to the experimenter(s)?

3.21 Determine the treatment sum of squares for the following Latin square design.

A(3.7)	B(3.9)	C(6.4)	D(5.1)
B(4.0)	A(3.5)	D(5.1)	C(4.8)
C(6.1)	D(3.8)	A(3.3)	B(2.9)
D(4.1)	C(5.5)	B(4.7)	A(3.1)

Is there evidence that any of the assumptions were violated? Could the assumptions be formally tested for this number of observations and design layout? Explain.

3.22 Show that there are 12 possible 3×3 Latin squares.

3.23 Could a PBIB design be constructed for five treatments to be repeated four times with a block size of 4? Why, or why not? If not, indicate a change in the number of repeats and/or block size that would allow a PBIB design to be constructed.

3.24 Assume that the blocks that are used for one of the blocking factors in a Latin square design are selected at random from a population of blocks. Does that create a problem with the analysis of the data? Explain.

3.25 Consider the following example given by Natrella (1963, pp. 13–14). An experiment was performed to determine the effects of four different geometrical shapes of a certain film-type composition resistor on the current-noise of the resistors. A BIB design was used because only three resistors could be mounted on one plate. In the design layout below it can be observed that each pair of treatments occurs twice in a block, so $\lambda = 2$. The observations are logarithms of the noise measurements.

	Shapes (Treatments)			
Plates (Blocks)	A	B	C	D
1	1.11		0.95	0.82
2	1.70	1.22		0.97
3	1.66	1.11	1.52	
4		1.22	1.54	1.18

Analyze the data by performing the intrablock analysis (using either hand computation or software) and draw appropriate conclusions.

3.26 The following ANOVA table appears in "Induced resistance in agricultural crops: Effects of jasmonic acid on herbivory and yield in tomato plants" by J. S. Thayer (*Environmental Entomology*, 28(1), 30–37). The response variable was the level of the enzyme polyphenol oxidase 8 d.

Source	MS	F	df	p
Treatment	69.68	14.51	1	0.001
Block	149.42	31.11	1	<0.001
Treatment X Block	5.68	1.18	1	0.29
Error	4.80		26	

(a) Is there anything odd or incorrect about this table, relative to an RCB design? If so, what is it? In particular, explain why this table could not be right if these are the only sources of variation and these sources are not a subset of a larger set of sources (which would be misleading).

(b) How many treatments and how many blocks were used, and what was the block size?

(c) The raw data were not given by the author, which precludes a full analysis of the data. Does the table above suggest that it was appropriate to use the RCB design? Explain.

3.27 Mead, Curnow, and Hasted (1993, references) gave an example of an RCB design for a study that involved three drugs and a placebo, with the experimental units being mice—four from each of five litters for a total of 20 mice. The objective was to determine if the drugs affected lymphocyte production. The data are as follows, with the letters A, B, C, and D representing the three drugs plus the placebo, respectively.

		Litter				
		1	2	3	4	5
Mouse within Litter	1	6.7 (B)	5.4 (D)	6.2 (C)	5.1(B)	6.2 (C)
	2	7.1 (C)	6.1 (A)	5.9 (B)	5.2 (D)	5.8 (B)
	3	6.7 (D)	5.8 (C)	6.9 (A)	5.0 (C)	5.3 (D)
	4	7.1 (A)	5.1 (B)	5.7 (D)	5.6 (A)	6.4 (A)

Perform an appropriate analysis, treating the drugs as fixed and the litters as random. Could both Treatment $\times$ Block and Error be shown in the ANOVA table, as in the preceding problem? Why, or why not?

3.28 Construct an example of an RCB design with five blocks and four levels of the factor such that the F-statistic for testing the factor effect is zero. Then construct an example (possibly different) such that both the F-statistic for testing the factor effect and the F-statistic for testing the block effect are zero.

3.29 Analyze the data given in Exercise 3.25 using ANOM.

3.30 Nyachoti, House, Slominski, and Seddon (2005, references) used a replicated 3×3 Latin square to determine the digestibilities of nutrients that comprise different diets. There were thus three diets and "... the pigs were randomly divided into two groups of three pigs each and within groups assigned to the experimental groups in a 3×3 Latin square design to give six observations per diet." Would you have designed this experiment differently if considerably more than six pigs were available for the experiment? If so, how would you have proceeded? If not, explain why not. In particular, do you believe that having multiple observations per pig could cause a problem, even though "period" is one of the blocking variables? Why, or why not?

3.31 Sarrazin, Mustafa, Chouinard, and Sotocinal (2004, references) studied the effects of dietary treatments on yield and composition of milk by using nine Holstein cows in three 3×3 Latin squares. These were not three replications, however, as different effects were studied in each Latin square, although two Latin square replications were used in one of the studies. In one of the other studies, "three multiparous lactating Holstein cows ... were used in a 3×3 Latin square design ..." This meant that there were repeated measures, with the different sampling times constituting the repeated measures. Read the article, if possible. Would you have used something other than Latin square designs since the overall study was multifaceted? Explain. Since repeated measures were involved, this problem is revisited in Exercise 8.7, in Chapter 8, which covers repeated measures designs.