Chapter 9
Multi-Factor Between-Subjects Designs

9.1 OVERVIEW

In this chapter, we consider between-subjects designs that include two or more factors. In multi-factor designs, the experimental conditions are formed by creating every combination of levels of the independent variables. There are two advantages of such designs. (1) The obvious advantage is economy; the effects of each of several factors can be studied in the same experiment. (2) The second advantage is that the combined, or *interaction*, effects of several variables can be studied. A specific example may help illustrate these two points. Wiley and Voss (1999) had students read about the Irish potato famine of the first half of the nineteenth century. One factor was the format: whether the material was presented in a single, textbook-like chapter (text format) or divided among eight sources in a computer, web-like (web format) environment. The second factor was the instruction subjects received; they were told to write either a narrative (N), a summary (S), an explanation (E), or an argument (A) about what produced changes in Ireland's population between 1800 and 1850. Thus, there were eight conditions in the experiment corresponding to the two formats combined with the four types of instruction. Following reading, students were tested on the material.

Wiley and Voss had several hypotheses that could be tested in their design. One hypothesis was that the more difficult web format would lead to a deeper understanding by forcing readers to integrate material obtained from several sources. Therefore, one question was whether the formats differed in their average effects. A second hypothesis was that the argument instruction would promote "more conceptual understanding." Therefore, a second question was whether the instructions would differ in their average effect. Finally, a third question was whether the size of any difference in performance between the argument instruction and the other instructions would depend upon the format. This is a question of whether there is an *interaction* between format and instructions.

These questions point to the two main goals of this chapter:

- To extend the models, assumptions, and analyses of Chapter 8 to deal with multi-factor between-subjects designs.

- To introduce the concept of interaction, providing illustrations of its analysis and interpretation.

9.2 THE TWO-FACTOR DESIGN: THE STRUCTURAL MODEL

9.2.1 The Model Equation

In these designs, there are two independent variables, A and B, with a levels of A, b levels of B, and n scores in each of the ab cells. A score will be represented as Y_{ijk}, the i^{th} score at the j^{th} level of A and the k^{th} level of B. We want to test hypotheses about population means and therefore need a model that relates the observed scores to the means of the populations formed by the combinations of the variables A and B, and to the error component of each score. These population parameters and the sample statistics that estimate them are presented in Table 9.1.

Consider a specific combination of levels of A and B, say, A_j and B_k. In terms of the *Wiley–Voss* experiment described in Section 9.1, this would be a cell formed by the combination of one of the two formats and one of the four instructions. Because the same combination of treatments is applied to everyone in that cell, the scores in the population corresponding to A_j and B_k should vary only because of error variance due to individual differences in factors such as ability, motivation, or physical state, or differences in other factors that can affect performance. Stating this more formally,

$$Y_{ijk} = \mu_{jk} + \varepsilon_{ijk} \tag{9.1}$$

where μ_{jk} is defined in Table 9.1 as the mean of the population formed by A_j and B_k, and ε_{ijk} is the error component of the i^{th} score in the cell formed by A_j and B_k in the experiment.

Now consider the possibility that scores in different combinations of treatments may differ systematically. It is useful to express the deviation of a score from the grand mean of all the populations by subtracting μ from both sides of Equation 9.1:

$$Y_{ijk} - \mu = (\mu_{jk} - \mu) + \varepsilon_{ijk} \tag{9.2}$$

The μ_{jk} may vary for any of three reasons; namely, they correspond to different levels of A and B,

Table 9.1 Population parameters and estimates for a two-factor design

Population parameters	Estimates
μ_{jk} = mean of the population of scores at A_j and B_k	$\bar{Y}_{.jk} = \sum_i Y_{ijk}/n$
$\mu_{j.} = \sum_k \mu_{jk}/b$ = mean of the populations in condition A_j	$\bar{Y}_{j..} = \sum_i \sum_k Y_{ijk}/nb$
$\mu_{.k} = \sum_j \mu_{jk}/a$ = mean of the populations in condition B_k	$\bar{Y}_{..k} = \sum_i \sum_j Y_{ijk}/na$
$\mu = \sum_j \sum_k \mu_{jk}/ab$ = mean of all ab populations	$\bar{Y}_{...} = \sum_i \sum_j \sum_k Y_{ijk}/nab$
$\varepsilon_{ijk} = Y_{ijk} - \mu_{jk}$ = error component of Y_{ijk}	$e_{ijk} = Y_{ijk} - \bar{Y}_{.jk}$

and to different combinations of those levels. We can represent this idea by the following identity:

$$(\mu_{jk} - \mu) = (\mu_j - \mu) + (\mu_k - \mu) + [(\mu_{jk} - \mu) - (\mu_j - \mu) - (\mu_k - \mu)] \tag{9.3}$$

A simpler notation is

$$\mu_{jk} - \mu = a_j + \beta_k + (\alpha\beta)_{jk} \tag{9.4}$$

Equation 9.4 states that $\mu_{jk} - \mu$ is a sum of three effects. The first of these, α_j, is the *main effect* of treatment A; its value represents the extent to which the average score in the population defined by the treatment A_j differs from the mean of all the scores in the ab populations. Similarly, β_k is the main effect of treatment B and reflects the extent to which the average score in population B differs from the average of all the scores in the ab populations. Finally, $(\alpha\beta)_{jk}$ is the *interaction effect* of A and B. The interaction is the difference between μ_{jk} and μ that remains after removal of the main effects of A and B; that is,

$$(\alpha\beta)_{jk} = (\mu_{jk} - \mu) - (\mu_j - \mu) - (\mu_k - \mu) \tag{9.5a}$$

This interaction effect is more often represented by simplifying the right-hand term in Equation 9.5a:

$$(\alpha\beta)_{jk} = \mu_{jk} - \mu_j - \mu_k + \mu \tag{9.5b}$$

From Equation 9.4, we can substitute for $\mu_{jk} - \mu$ in Equation 9.2. Recognizing that nuisance variables will also contribute to the observed data, we have the structural model underlying tests of hypotheses for the two-factor design:

$$Y_{ijk} - \mu = \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk} \tag{9.6}$$

In words, the variability among scores in the populations has four possible sources: the effects of manipulating A; the effects of manipulating B; the joint, or interaction, effects of A and B; and the error component. A more detailed summary of the model components and related assumptions is presented in Box 9.1.

Box 9.1 Components of the Structural Model

1. *The error component,* $\varepsilon_{ijk} = Y_{ijk} - \mu_{jk}$. The errors are independently and normally distributed with mean zero and variance σ^2_e, within each treatment population defined by a combination of levels of A and B.
2. *The main effect of treatment* A, $\alpha_j = \mu_j - \mu$. The factor A is assumed to have fixed effects; that is, the a levels have been arbitrarily selected and are viewed as representing the population of levels. Then $\Sigma_j \alpha_j = 0$. The F test of the A main effect tests the null hypothesis that

   $$H_0: \mu_{1.} = \mu_{2.} = \ldots = \mu_{j.} = \ldots = \mu_{a.}$$

   or, equivalently,

   $$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_j = \ldots = \alpha_a = 0$$

3. *The main effect of treatment* B, $\beta_k = \mu_k - \mu$. This is also a fixed-effect variable and so $\Sigma_k \beta_k = 0$. The F test of the B main effect tests the null hypothesis that

   $$H_0: \mu_{.1} = \mu_{.2} = \ldots = \mu_{.k} = \ldots = \mu_{.b}$$

   or, equivalently,

   $$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = \ldots = \beta_b = 0$$

4. The *interaction effect of* A_j *and* B_k, $(\alpha\beta)_{jk} = \mu_{jk} - \mu_j - \mu_k + \mu$. Because both A and B have fixed effects, $\Sigma_j(\alpha\beta)_{jk} = \Sigma_k(\alpha\beta)_{jk} = 0$. The relevant null hypothesis is

$$H_0: (\alpha\beta)_{11} = (\alpha\beta)_{12} = \ldots = (\alpha\beta)_{jk} = \ldots = (\alpha\beta)_{ab} = 0$$

Equation 9.6 provides the basis for the analysis of variance for the two-factor design. We will develop this relation between the analysis of variance and the structural model in Section 9.3. However, before doing so, we will try to provide a better understanding of what the main and interaction effects represent.

9.2.2 Understanding Main and Interaction Effects

Assume that we have a 2×4 design. Further assume that the population means are those in Panel *a* of Table 9.2. Note that the main effects (α_j and β_k) are calculated by subtracting the *grand mean* (obtained by averaging over all scores) from the *marginal means* (those obtained by averaging all the means in a row or column). For designs with equal n, the main effects are independent or

Table 9.2 Treatment population means and effects

(*a*) Original population means

	B_1	B_2	B_3	B_4	μ_j	$\alpha_j = \mu_j - \mu$
A_1	65	50	47	58	55	5
A_2	43	48	51	38	45	−5
μ_k	54	49	49	48	$\mu = 50$	
$\beta_k = \mu_k - \mu$	4	−1	−1	−2		

(*b*) Population means after removing the A (α_j) and B (β_k) main effects

	B_1	B_2	B_3	B_4	Mean
A_1	56	46	43	55	50
A_2	44	54	57	45	50
Mean	50	50	50	50	$\mu = 50$

(*c*) Interaction effects; $(\alpha\beta)_{jk} = (\mu_{jk} - \mu) - \alpha_j - \beta_k$

	B_1	B_2	B_3	B_4	Mean
A_1	6	−4	−7	5	0
A_2	−6	4	7	−5	0
Mean	0	0	0	0	

orthogonal. That is, knowing how the means of one factor vary across levels tells us nothing about how the means vary on the other factor.

In Panel *b*, we have the population means after the main effects have been subtracted. For example, in the A_1B_2 population, the cell mean after removal of the main effects that have contributed to it is $50 - 5 - (-1) = 46$. Note that although the marginal means are now identical, the adjusted cell means still vary. The reason for this variation is the presence of interaction effects. If we subtract the grand mean, μ, from each of the values in Panel *b*, we obtain the results in Panel *c* of Table 9.2; these are the interaction effects associated with each cell. In summary, one definition of interaction effects is that they are the difference between the cell mean and the grand mean, after removing the main effects of the independent variables.

It is useful to compare the pattern of means in Panel *a* of Table 9.2 with the pattern in Panel *a* of Table 9.3. The means in Table 9.3 show exactly the same main effects of *A* and *B* as the means in Table 9.2, as seen by comparing the corresponding values of α_j and β_k of the two tables. However, Table 9.3 does not present an interaction; in that case, we say that *A* and *B* have *additive effects* because the mean of each cell is determined by *adding* the main effect of *A* and the main effect of *B* to the grand mean. The absence of interaction effects is shown by subtracting the row and column effects from each mean. For example, in the A_1B_1 cell, $59 - 5 - 4 = 50$; doing the same for all cells, the values are all also 50, as shown in Panel *b*. An important point that is implicit in this comparison of Tables 9.2 and 9.3 is that the magnitudes of the main effects and interaction are unrelated, or *orthogonal* to one another, in a design where *n* is equal for all conditions. Thus, the presence of one or both main effects does not tell us anything about the magnitude of the interaction, and vice versa.

Another way of comparing the two tables is particularly useful for understanding the meaning of an interaction. In Panel *a* of Table 9.4, the effect of *A* is computed at each level of *B* for the data of Table 9.2; this is done by subtracting the A_2 mean from the A_1 mean. These *simple effects* of *A* are also computed in Panel *b* for the data of Table 9.3. The key observation is that the values of the simple effects of *A* differ over levels of *B* in Panel *a*, indicating that *A* and *B* interact. In contrast, the

Table 9.3 Treatment population means with no interaction effects

	B_1	B_2	B_3	B_4	μ_j	$\alpha_j = \mu_j - \mu$
A_1	59	54	54	53	55	5
A_2	49	44	44	43	45	−5
μ_k	54	49	49	48	$\mu = 50$	
$\beta_k = \mu_k - \mu$	4	−1	−1	−2		

(*b*) Population means after removing the *A* (α_j) and *B* (β_k) main effects

	B_1	B_2	B_3	B_4	Mean
A_1	50	50	50	50	50
A_2	50	50	50	50	50
Mean	50	50	50	50	$\mu = 50$

Table 9.4 Simple effects of A at each level of B ($\mu_{1k} - \mu_{2k}$) for the data of Tables 9.2 and 9.3

(a) Population means from Table 9.2 with interaction effects

	B_1	B_2	B_3	B_4	μ_i
A_1	65	50	47	58	55
A_2	43	48	51	38	45
$\mu_{1k} - \mu_{2k}$	22	2	−4	20	10

(b) Population means from Table 9.3 with no interaction effects

	B_1	B_2	B_3	B_4	μ_i
A_1	59	54	53	53	55
A_2	49	44	43	43	45
$\mu_{1k} - \mu_{2k}$	10	10	10	10	10

simple effects of A are constant over levels of B in Panel b, indicating no interaction. Thus, *an interaction means that the size of the effect of factor A differs over levels of factor B* or, equivalently, that the effect of factor B depends on the level of factor A. The same observation may be made graphically. The means for Panel a of Table 9.4 are graphed in Panel a of Fig. 9.1; the means for Panel b of Table 9.4 are graphed in Panel b of Fig. 9.1. The obvious difference in the two graphs is that the curves in Panel a are not parallel, whereas the curves in Panel b are. Thus, with respect to the population means, an *interaction is a departure from parallelism*. We say that the interaction is a significant source of variance when the size of the effects of one variable changes significantly across levels of the other variable.

9.3 TWO-FACTOR DESIGNS: THE ANALYSIS OF VARIANCE

The analysis of the variability in the data for the two-factor design follows the same logic developed for the one-factor design presented in Chapter 8. The structural model (Equation 9.6) suggests a way to partition the deviation of a score from the grand mean, $Y_{ijk} - \bar{Y}_{...}$, into several components. We derive these components in Section 9.3.1, and then develop formulas for the sums of squares based on them in Section 9.3.2. We then construct mean squares and, subsequently, F statistics to test null hypotheses about main and interaction effects.

9.3.1 Components of the Scores

In Section 9.2.1, we assumed the structural model

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk}$$

After subtracting μ from both sides and substituting treatment population means, we have

$$Y_{ijk} - \mu = (\mu_j - \mu) + (\mu_k - \mu) + (\mu_{jk} - \mu_j - \mu_k + \mu) + \varepsilon_{ijk} \tag{9.7}$$

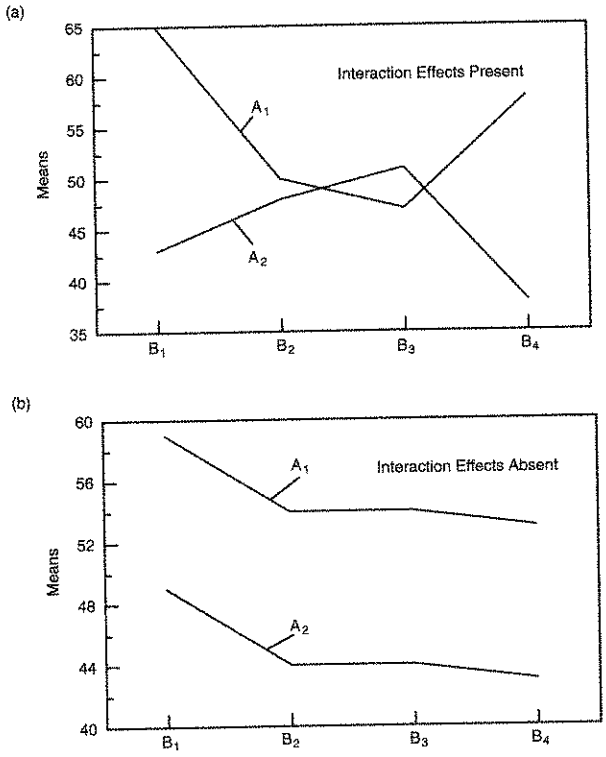

Fig. 9.1 Plots of the population means in Tables 9.2 and 9.3.

Substituting the estimates of these parameters from Table 9.1, we have the basis for the ANOVA:

$$(Y_{ijk} - \overline{Y}_{...}) = (\overline{Y}_{j.} - \overline{Y}_{...}) + (\overline{Y}_{.k} - \overline{Y}_{...}) + (\overline{Y}_{jk} - \overline{Y}_{j.} - \overline{Y}_{.k} + \overline{Y}_{...}) + (Y_{ijk} - \overline{Y}_{jk}) \tag{9.8}$$

In words,

score − grand mean = main effect of A + main effect of B + interaction effect + residual error

9.3.2 Sums of Squares

Equation 9.8 forms the basis for the sums of squares; these are important components in tests of null hypotheses. Squaring both sides of Equation 9.8 and summing yields the sums of squares (SS) formulas in Panel *a* of Table 9.5. As a first step, we partitioned the total sum of squares into two components: a between-cells sum of squares and a within-cell (S/AB) sum of squares. The between-cell variability usually is not of interest in itself because it has several possible sources. For example, in the *Wiley–Voss* data set, the eight cell means may differ because they represent different formats, different instructions, or different combinations of formats and instructions. Although software packages do not include the between-subjects variability, SS_{cells}, we include it because it is involved in our conceptualization and calculation of the SS_{AB} and the $SS_{S/AB}$. The components of the between-cells sum of squares in Table 9.5 correspond to the main and interaction sources of variance that we wish to test.

The formulas presented in Table 9.5 define the sums of squares. Although statistical software

Table 9.5 The analysis of variance (ANOVA) table for the two-factor between-subjects design (*a*) and expected mean squares (*b*)

(*a*) ANOVA

Source	*df*	*SS*	*MS*	*F*
Total	$abn - 1$	$\sum_{j}^{a}\sum_{k}^{b}\sum_{i}^{n}(Y_{ijk} - \bar{Y}_{...})^2$		
Between cells	$ab - 1$	$n\sum_{j}^{a}\sum_{k}^{b}(\bar{Y}_{jk} - \bar{Y}_{...})^2$		
A	$a - 1$	$nb\sum_{j}^{a}(\bar{Y}_{j..} - \bar{Y}_{...})^2$	SS_A/df_A	$MS_A/MS_{S/AB}$
B	$b - 1$	$na\sum_{k}^{b}(\bar{Y}_{.k.} - \bar{Y}_{...})^2$	SS_B/df_B	$MS_B/MS_{S/AB}$
AB	$(a - 1)(b - 1)$	$SS_{cells} - SS_A - SS_B$	SS_{AB}/df_{AB}	$MS_{AB}/MS_{S/AB}$
S/AB	$ab(n - 1)$	$SS_{total} - SS_{cells}$	$SS_{S/AB}/df_{S/AB}$	

(*b*) Expected mean squares

SV	*EMS*
A	$\sigma_e^2 + nb\sum_{j}\alpha_j^2/(a-1) = \sigma_e^2 + nb\theta_A^2$
B	$\sigma_e^2 + na\sum_{k}\beta_k^2/(b-1) = \sigma_e^2 + na\theta_B^2$
AB	$\sigma_e^2 + n\sum_{j}\sum_{k}(\alpha\beta)_{jk}^2/(a-1)(b-1) = \sigma_e^2 + n\theta_{AB}^2$
S/AB	σ_e^2

Note: $\alpha_j = \mu_{j.} - \mu$, $\beta_k = \mu_{.k} - \mu$, and $(\alpha\beta)_{jk} = (\mu_{jk} - \mu) - \alpha_j - \beta_k = (\mu_{jk} - \mu_{j.} - \mu_{.k} + \mu)$. The θ^2 notation serves as a reminder that $\Sigma\alpha_j^2/(a-1)$ is not a variance; the variance of the treatment population means has *a* as the denominator.

usually will be available to perform the calculations, the formulas are presented to remind us that the sums of squares are indices of variability. For example, the SS_{total} is $abn - 1$ times the variance of all the scores, the SS_A is $bn(a-1)$ times the variance of the A marginal means, and the SS_{cells} is $n(ab-1)$ times the variance of the ab cell means. The tests of null hypotheses corresponding to the A, B, and AB sources of variance (SV) test whether those variances are greater than chance.

9.3.3 Degrees of Freedom

A formula for degrees of freedom is associated with each of the sources of variances. The df_{total} are $abn - 1$ because this SV represents the variability of all abn scores about the grand mean. The SS_{cells} is distributed on $ab - 1$ df because it involves the variability of the ab cell means about the grand mean. The df for the main effects have the same form as in Chapter 8; these SV reflect the variance

of the a (or b) means about the grand mean and therefore one df is lost. The interaction degrees of freedom are

$$df_{AB} = (ab - 1) - (a - 1) - (b - 1) = (a - 1)(b - 1)$$

reflecting the adjustment of cell variability for the variability due to A and B. In practice, we can generate the degrees of freedom for an interaction just by multiplying the degrees of freedom for the interacting variables.

The $df_{S/AB}$ may be thought of as the difference between df_{total} and df_{cells}:

$$df_{S/AB} = (abn - 1) - (ab - 1)$$

These degrees of freedom may also be viewed as the result of summing the degrees of freedom for variability within each cell; there are ab cells, each with $n - 1$ df, yielding $ab(n - 1)$ df. The two ways of thinking about degrees of freedom, as a difference between the total df and the cell df, or as a sum over cells, are equivalent: $(abn - 1) - (ab - 1) = ab(n - 1)$.

9.3.4 Mean Squares (*MS*), Expected Mean Squares (*EMS*), and *F* Ratios

As in the one-factor design, the MS of Table 9.5 are ratios of SS to df. Conceptually, however, the mean squares for the main effects are simple functions of variances. For example, MS_A is the variance of the a marginal means in the A conditions, multiplied by nb, the number of scores upon which each mean is based. Similarly, MS_B is the variance of the b marginal means in the B conditions, multiplied by na, the number of scores upon which each mean is based. The error mean square, $MS_{S/AB}$, is an average of the within-cell variances; it can be calculated as

$$MS_{S/AB} = (1/ab) \sum_j \sum_k s^2_{jk}$$

where s^2_{jk} is the variance of the n scores in the cell defined by A_j and B_k.[1]

All three F ratios are formed by dividing by $MS_{S/AB}$. This is justified by the expected mean squares (EMS; see Panel b of Table 9.5). As we stated in Chapter 8, forming a ratio of two mean squares follows the rule that the numerator and denominator MS must have the same expectation when the null hypothesis corresponding to the numerator is true.

The EMS are derived by assuming the structural model of Equation 9.6, independence of the scores, and homogeneity of the population variances. In addition, if the treatment populations are normally distributed and the null hypothesis is true, the ratio of mean squares is distributed as F.[2]

We now have both a conceptual framework and formulas on which to base tests of hypotheses about main and interaction effects. We next apply this framework to the analysis of the inference verification test (IVT) data in Table 9.6. The complete data set, reported by Wiley and Voss (1999), includes a number of other measures and may be found in the *Wiley* file among the *Data Sets* pages on the website for this book.

[1] Note that the mean square for the interaction is related to the variance of the cell means, but not as simply as is the case for the main effects. The variability associated with both main effects must be removed from the variability in the cell means; this happens at the level of the sums of squares calculations.

[2] Strictly speaking, if the null hypothesis is true, we have the central F distribution whose cutoffs are presented in Appendix Table C.5. However, if the null hypothesis is not true, we have members of the noncentral F distribution family. The noncentral F distribution can be used to perform power calculations.

Table 9.6 Inference scores (percent correct) from Wiley and Voss (1999) with summary statistics

	Instructions[a]				
Format	*N*	*S*	*E*	*A*	
Text	70	50	70	70	
	80	90	80	70	
	80	60	70	60	
	70	80	60	60	
	60	70	60	70	
	50	80	80	90	
	80	80	70	90	
	80	70	60	80	
$\bar{Y}_{Text,k} =$	71.25	72.5	68.75	73.75	$\bar{Y}_{Text} = 71.56$
$s^2_{Text,k} =$	126.79	164.29	69.64	141.07	
Web	100	70	60	100	
	80	70	60	90	
	60	80	80	100	
	60	50	80	80	
	60	90	80	90	
	70	60	60	100	
	90	100	80	70	
	90	70	80	90	
$\bar{Y}_{Web,k} =$	76.25	73.75	72.5	90	$\bar{Y}_{Web} = 78.1$
$s^2_{Web,k} =$	255.36	255.36	107.14	114.29	
$\bar{Y}_{j,Instruct} =$	73.75	73.13	70.63	81.88	$\bar{Y}_{..} = 74.84$

[a] *N* = Narrative, *S* = Summary, *E* = Explanation, and *A* = Argument.

9.3.5 The *Wiley–Voss* Example

Before considering the ANOVA, we should get some sense of the effects of our variables. Looking at the two marginal format means in the right-most column of Table 9.6, we find that performance for the web format ($\bar{Y}_{Web}$) was better than that for the text format ($\bar{Y}_{Text}$). This difference between the web and text formats is largely due to the argument (*A*) instructional condition; although the web format has a higher mean than the text format in all instructional conditions, the differences between web and text cell means are small except in the *A* column. Turning next to the marginal instructional means ($\bar{Y}_N$, $\bar{Y}_S$, $\bar{Y}_E$, and $\bar{Y}_A$), we find the *IVT* mean to be higher in the argument condition than in any of the others. Again, however, we must qualify this; the advantage of the argument condition is quite pronounced for the web format, but rather small in the text format. Whether we view the data as showing that the difference between format means depends on instructions, or as showing that the differences among instructional means depend on format, our focus should be on the interaction of format and instructions. This is clearer in the bar graph of Fig. 9.2. Although web learning has an advantage in all four instructional conditions, that advantage is clearly larger in the *A* condition than in any of the other three.

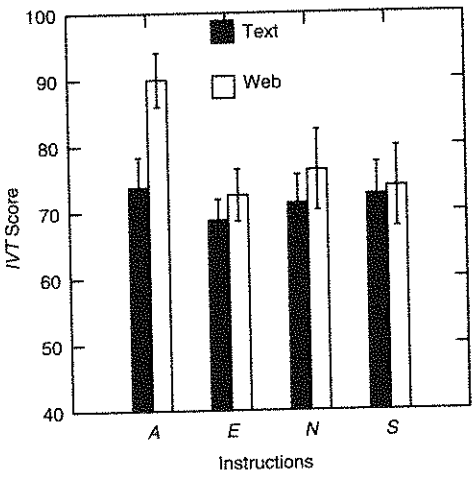

Fig. 9.2 Bar graph of the Wiley–Voss (1999) *IVT* data.

In addition to examining the means, we also checked for any departures from the assumptions underlying the hypothesis tests we wished to perform. Box plots and the Levene tests failed to reveal any violations of the assumption of homogeneity of variance severe enough to undermine conclusions based on the tests of main and interaction effects. Nor were any outliers present in the box plots. Finally, plots of residuals (deviations of scores from the cell means) and significance tests revealed no departure from normality. In summary, the ANOVA appears to provide appropriate tests for the *IVT* data.

The results of the analysis of variance for the data are presented in Table 9.7. As the experimenters hypothesized, a significantly higher proportion of inferences were correctly verified by subjects in the web than in the text format condition. This means that the average of the four populations of *IVT* scores obtained under the web format differs from the average of the four populations of *IVT* scores obtained under the text format. In terms of Table 9.6, it is the *marginal means*, $\bar{Y}_{Web}$ (78.13) and $\bar{Y}_{Text}$ (71.56), that differ significantly. We can conclude that, *averaging over levels of instruction*, the effects of the two formats differ significantly. However, this does not provide information about the effects of the formats at any particular level of instructions.

The experimenters also were interested in whether instructions would affect performance. They reported that the effect was "marginally significant" because the p-value was .07, short of the .05 level usually required for statistical significance. We will consider further the question of the effect

Table 9.7 The analysis of variance (ANOVA) table for the Wiley–Voss data

SV	*df*	*SS*	*MS*	*F*	*p*
Total	63	10,998.44			
Between cells	7	2,360.94			
Format (F)	1	689.06	689.06	4.47	.039
Instructions (I)	3	1,142.19	380.73	2.47	.071
FI	3	529.69	176.56	1.14	.337
S/*FI*	56	8,367.50	154.24		

of instructions later in this chapter when we calculate various measures of effect size for the *Wiley–Voss* data.

The F test of the *Format* × *Instructions* interaction tests the null hypothesis that the effects of one variable are the same under all levels of the other variable. One statement of the null hypothesis of no interaction is that the difference between the text and web population means is the same under all types of instructions. This may be represented as

$$H_0\colon (\mu_{Text,\,N} - \mu_{Web,\,N}) = (\mu_{Text,\,S} - \mu_{Web,\,S}) = (\mu_{Text,\,E} - \mu_{Web,\,E}) = (\mu_{Text,\,A} - \mu_{Web,\,A})$$

where, for example, $\mu_{Text,\,N}$ is the mean of the population of scores obtained under the text format and the narrative instructions.

An equivalent statement of the null hypothesis of no interaction is that the effects of instructions are the same at the two format levels. We could state this null hypothesis as

$$H_0\colon (\mu_{Text,\,N} - \mu_{Text,\,S}) = (\mu_{Web,\,N} - \mu_{Web,\,S})$$

and

$$(\mu_{Text,\,S} - \mu_{Text,\,E}) = (\mu_{Web,\,S} - \mu_{Web,\,E})$$

and

$$(\mu_{Text,\,E} - \mu_{Text,\,A}) = (\mu_{Web,\,E} - \mu_{Web,\,A})$$

Both forms of H_0 are ways of stating that the effect of one factor is the same at each level of the other factor (i.e., parallel functions).

The difference between the observed text and web means under argument (A) instructions appears considerably larger than the other differences, as evidenced in the means of Table 9.6 and the bar graph of Fig. 9.2. Nevertheless, the F test of the interaction fell well short of significance. This raises several questions. Given that the *Instructions* and the *Format* × *Instructions* sources of variance were not significant, is it proper to test more specific hypotheses related to those sources such as whether the narrative and argument means differ? If so, should we have different criteria for significance than for the usual t test? And what should our error term be? We will consider those questions, and attempt to further clarify the concept of interaction, in Chapter 10.

In sum, the analyses of the *Wiley–Voss* data are somewhat ambiguous. The effect of instructions is marginally significant. Graphically, there appears to be an interaction, but the F test does not confirm its reliability. The only significant result is the effect of format, showing better performance with the web presentation. However, there is some doubt about how to interpret this seemingly straightforward result because of the question about whether the effect of format is due to the results from the Argument condition.

At this point, we complete our presentation of the basic ANOVA by extending it to designs with more than two factors.

9.4 THREE-FACTOR BETWEEN-SUBJECTS DESIGNS

9.4.1 The General Case

Extending the two-factor design to the three-factor design is straightforward; the only new concept is the three-factor interaction. Therefore, we will present the general case of the three-factor design concisely so that we may reinforce the basic concepts already developed for the simpler one-factor and two-factor designs.

The general case of the three-factor between-subjects design involves a levels of A, b levels of B, c levels of C, and n scores in each of the abc cells. The relevant indices are

$$i = 1, 2, \ldots, n; j = 1, 2, \ldots, a; k = 1, 2, \ldots, b; \text{ and } m = 1, 2, \ldots, c$$

The structural model looks much like that for the two-factor experiment except that there are now three two-factor interactions and there is the added three-factor interaction:

$$Y_{ijkm} = \mu + \alpha_j + \beta_k + \gamma_m + (\alpha\beta)_{jk} + (\alpha\gamma)_{jm} + (\beta\gamma)_{km} + (\alpha\beta\gamma)_{jkm} + \varepsilon_{ijkm} \tag{9.9}$$

Definitions in terms of population means, together with estimates of those means, are presented in Table 9.8. The only new definition is that of the interaction effect for the cell $A_jB_kC_m$; this is the difference between the cell mean and the grand mean, adjusted for all main and first-order interaction effects that contribute to the cell.

The sums of squares follow directly from the parameter estimates by squaring each term in the Estimate column of Table 9.8, and summing over the indices. The results of this process are presented in Table 9.9 together with the degrees of freedom. The only new df term is $(a-1)(b-1)(c-1)$, the df_{ABC}. This follows by subtracting the main and two-factor df from $abc - 1$, the between-cells df.

The mean squares are obtained, as usual, by dividing sums of squares by degrees of freedom. $MS_{S|ABC}$, the average within-cell variance, is the error term against which all main and interaction terms are tested. As in other designs, expected mean squares, presented in Table 9.10, provide the rationale for this choice of error terms. These have been derived from the structural model (Equation 9.9) under the usual assumptions that the scores in the abc treatment populations are independently distributed and that the population variances all equal σ_ε^2. In addition, all three factors are assumed to have *fixed effects*; that is, the levels have been arbitrarily selected and not randomly sampled from a universe of treatment levels.

9.4.2 Extending the *Wiley–Voss* Example

To illustrate the concepts and analyses, we add a hypothetical third factor to the design of the *Wiley–Voss* experiment. Assume that there are only two levels of instruction (I), Summary and Argument, and two formats (F), Text and Web. Further assume that subjects are divided with respect to experience (E)—either those who had prior experience searching the Internet (experts) or

Table 9.8 Parameters of the structural model for a three-factor design

Source	Population parameter	Estimate
A	$\alpha_j = \mu_{j..} - \mu_{...}$	$\overline{Y}_{j..} - \overline{Y}_{...}$
B	$\beta_k = \mu_{.k.} - \mu_{...}$	$\overline{Y}_{.k.} - \overline{Y}_{...}$
C	$\gamma_m = \mu_{..m} - \mu_{...}$	$\overline{Y}_{..m} - \overline{Y}_{...}$
AB	$(\alpha\beta)_{jk} = (\mu_{jk.} - \mu_{...}) - \alpha_j - \beta_k$	$\overline{Y}_{jk.} - \overline{Y}_{j..} - \overline{Y}_{.k.} + \overline{Y}_{...}$
AC	$(\alpha\gamma)_{jm} = (\mu_{j.m} - \mu_{...}) - \alpha_j - \gamma_m$	$\overline{Y}_{j.m} - \overline{Y}_{j..} - \overline{Y}_{..m} + \overline{Y}_{...}$
BC	$(\beta\gamma)_{.km} = (\mu_{.km} - \mu_{...}) - \beta_k - \gamma_m$	$\overline{Y}_{.km} - \overline{Y}_{.k.} - \overline{Y}_{..m} + \overline{Y}_{...}$
ABC	$(\alpha\beta\gamma)_{jkm} = (\mu_{jkm} - \mu_{...}) - \alpha_j - \beta_k$ $- \gamma_m - (\alpha\beta)_{jk} - (\alpha\gamma)_{jm} - (\beta\gamma)_{km}$	$\overline{Y}_{jkm} + \overline{Y}_{j..} + \overline{Y}_{.k.} + \overline{Y}_{..m}$ $- \overline{Y}_{jk.} - \overline{Y}_{j.m} - \overline{Y}_{.km} - \overline{Y}_{...}$
Error	ε_{ijkm}	$Y_{ijkm} - \overline{Y}_{jkm}$

Table 9.9 Degrees of freedom and sums of squares in a three-factor design

Source	*df*	*SS*
Total	$abcn - 1$	$\sum_i \sum_j \sum_k \sum_m (Y_{ijkm} - \bar{Y}_{\ldots})^2$
Between cells	$abc - 1$	$n \sum_j \sum_k \sum_m (\bar{Y}_{\cdot jkm} - \bar{Y}_{\ldots})^2$
A	$a - 1$	$nbc \sum_j (\bar{Y}_{\cdot j \cdot \cdot} - \bar{Y}_{\ldots})^2$
B	$b - 1$	$nac \sum_k (\bar{Y}_{\cdot \cdot k \cdot} - \bar{Y}_{\ldots})^2$
C	$c - 1$	$nab \sum_m (\bar{Y}_{\cdot \cdot \cdot m} - \bar{Y}_{\ldots})^2$
AB	$(a - 1)(b - 1)$	$nc \sum_j \sum_k (\bar{Y}_{\cdot jk \cdot} - \bar{Y}_{\cdot j \cdot \cdot} - \bar{Y}_{\cdot \cdot k \cdot} + \bar{Y}_{\ldots})^2$
AC	$(a - 1)(c - 1)$	$nb \sum_j \sum_m (\bar{Y}_{\cdot j \cdot m} - \bar{Y}_{\cdot j \cdot \cdot} - \bar{Y}_{\cdot \cdot \cdot m} + \bar{Y}_{\ldots})^2$
BC	$(b - 1)(c - 1)$	$na \sum_k \sum_m (\bar{Y}_{\cdot \cdot km} - \bar{Y}_{\cdot \cdot k \cdot} - \bar{Y}_{\cdot \cdot \cdot m} + \bar{Y}_{\ldots})^2$
ABC	$(a - 1)(b - 1)(c - 1)$	$n \sum_j \sum_k \sum_m (\bar{Y}_{\cdot jkm} + \bar{Y}_{\cdot j \cdot \cdot} + \bar{Y}_{\cdot \cdot k \cdot} + \bar{Y}_{\cdot \cdot \cdot m} - \bar{Y}_{\cdot jk \cdot} - \bar{Y}_{\cdot j \cdot m} - \bar{Y}_{\cdot \cdot km} - \bar{Y}_{\ldots})^2$
S/ABC (within cells)	$abc(n - 1)$	$SS_{tot} - SS_{B\text{-}cells}$

those who were novices. Assume that there are 10 scores in each of the eight cells; i.e., $n = 10$ and $N = 80$. The means for this hypothetical experiment are presented in Table 9.11. The $MS_{S/cells}$ is presented as having a value of 154, but the relevant condition variances on which that value is based have been omitted. Note that we are dealing with the simplest case of a three-factor design, one in which there are only two levels of each of the three variables. The simplicity of the design enables us to concentrate on basic concepts. The design has the added advantage that it is a very common research design.

Main Effects. In the ANOVA, there will be three sources of main effects: instructions, format, and experience. We can view these main effects by calculating the marginal means separately for each factor. For example, the test of the format source of variance involves a comparison of the text and web marginal means. These means are obtained by averaging over the four combinations of instructions and experience. As can be seen in the right-hand column in the bottom panel of Table 9.11, the text and web means are 72.375 and 86.125. The significance test is a test of the null hypothesis that, *averaging over the four populations corresponding to the combinations of experience and instructions*, there is no difference between the population text and web means.

Table 9.10 Expected mean squares (*EMS*) for the three-factor design

SV	*EMS*
A	$\sigma_e^2 + nbc\sum_j \alpha_j^2/(a-1)$
B	$\sigma_e^2 + nac\sum_k \beta_k^2/(b-1)$
C	$\sigma_e^2 + nab\sum_m \gamma_m^2/(c-1)$
AB	$\sigma_e^2 + nc\sum_j\sum_k (\alpha\beta)_{jk}^2/(a-1)(b-1)$
AC	$\sigma_e^2 + nb\sum_j\sum_m (\alpha\gamma)_{jm}^2/(a-1)(c-1)$
BC	$\sigma_e^2 + na\sum_k\sum_m (\beta\gamma)_{km}^2/(b-1)(c-1)$
ABC	$\sigma_e^2 + n\sum_j\sum_k\sum_m (\alpha\beta\gamma)_{jkm}^2/(a-1)(b-1)(c-1)$
S/ABC	σ_e^2

Note: The parameters of the structural model are defined in Table 9.8.

Table 9.11 Means for a hypothetical extension of the Wiley–Voss experiment

		Summary	Argument	Mean
Novice	Text	71.25	73.75	72.50
	Web	76.50	90.00	83.25
	Mean	73.875	81.875	77.875
Expert	Text	70.50	74.00	72.25
	Web	88.25	89.75	89.00
	Mean	79.375	81.875	80.625
		Averaging over novices and experts		
	Text	70.875	73.875	72.375
	Web	82.375	89.875	86.125
	Mean	76.625	81.875	79.25

Note: Each cell contains 10 scores; i.e. $n = 10$, $N = 80$. The error mean square, $MS_{S/cells} = 154$.

To test this null hypothesis, we can calculate MS_F by computing the variance of the two marginal means, 94.53125, and then multiplying by the number of scores each mean is based on, 40, to obtain $MS_F = 3{,}781.25$. To construct the F test, divide MS_F by $MS_{S/cells} = 3{,}781.25/154 = 24.55$, which is significant on 1 and 72 degrees of freedom.

An astute reader might wonder whether the effect of format could be tested by a simple t test for independent means. In fact, we can calculate a t statistic:

$$t = \frac{\bar{Y}_{Web} - \bar{Y}_{Text}}{\sqrt{MS_{S/cells}\left(\frac{1}{n_{Web}} + \frac{1}{n_{Text}}\right)}}$$

$$= \frac{86.125 - 73.375}{\sqrt{154\left(\frac{1}{40} + \frac{1}{40}\right)}} = 4.59$$

Alternatively, noting that when the numerator has 1 degree of freedom, $F = t^2$; we could have calculated

$$F = \frac{(\bar{Y}_{Web} - \bar{Y}_{Text})^2 \Big/ \left(\frac{1}{n_{Web}} + \frac{1}{n_{Text}}\right)}{MS_{S/cells}} \tag{9.10}$$

The instruction and experience sources can be tested in a similar manner.[3] We encourage the reader to compute the mean squares for instruction and experience and confirm that they match the values in Table 9.12.

Table 9.12 The analysis of variance (ANOVA) table for the hypothetical data in Table 9.11

SV	*df*	*SS*	*MS*	*F*	*p*
Total	79				
Between cells	7	5,127.50	732.5		
Format (*F*)	1	3,781.25	3,781.25	24.55	.000
Instructions (*I*)	1	551.25	551.25	3.58	.05
Experience (*E*)	1	151.25	151.25	<1	
FI	1	101.25	101.25	<1	
FE	1	180.00	180.00	1.17	
IE	1	151.25	151.25	<1	
FIE	1	211.25	211.25	1.37	
S/FIE	72		154.00		

First-Order (Two-Factor) Interactions. There are three possible two-factor interactions in a three-factor design. In the *Wiley–Voss* example, they are: *Format* × *Instructions* (*FI*), *Format* × *Experience* (*FE*), and *Instructions* × *Experience* (*IE*). The *FE* interaction is of particular interest because computer experience should not have an effect in the text condition, but it well might in the

[3] Note that the t test computed in this example appropriately uses an error term based on the within-cell variances of the three-factor design. This is not the error term that would be calculated by a software program like SPSS if the user were to simply request a t test comparing the web and text formats. Rather, a simple t test that ignores the other two factors in the design would compute an error term that would be inflated by effects involving the two ignored factors.

web condition. The interpretation of this interaction is essentially the same as if F and E were the only factors in the experiment, except that in this case, the relevant means are obtained by averaging over the levels of the third variable, instructions. These means are in the right-most column in the upper two panels of Table 9.11. A better way to see the possible interaction is to redisplay the means:

		Experience	
		Expert	Novice
Format	Web	89.00	72.25
	Text	83.25	72.50

The pattern of means indicates that the advantage of the expert over the novice is, as hypothesized, greater when information is presented in the web format. The significance test of this interaction is a test of the null hypothesis that, *averaging over the two levels of instructions*, there is no difference between experts and novices in the magnitude of the effect of format.

Calculating the interaction sum of squares is generally too tedious to do by hand. However, in the case of interactions based on 1 *df*, we can represent the interaction as a difference between differences. This makes hand calculations more manageable; much more importantly, expressing a 2×2 interaction as a difference between simple effects is a better way to understand what is being tested. Let I indicate the interaction effect; then,

$$I_{FE} = (89.00 - 83.25) - (72.25 - 72.50) = 6.00$$

In words, the effect of format is 6 points greater for experts than for novices. Rearranging terms, this FE interaction can be rewritten as a difference between the sums of diagonal cell means:

$$I_{FE} = (89.00 + 72.50) - (83.25 + 72.25)$$

If we calculate the averages of the two diagonal sums, we have the basis for a test of the difference between two means; these are $(89.00 + 72.50)/2$, or 80.75, and $(83.25 + 72.25)/2$, or 77.75. As with our example in testing the main effect of format, we can test the FE interaction by comparing these two means via a t test:

$$t = \frac{80.75 - 77.75}{\sqrt{154\left(\frac{1}{40} + \frac{1}{40}\right)}} = 1.081$$

Alternatively, applying Equation 9.10, we can construct an F test:

$$F = \frac{(80.75 - 77.75)^2 \Big/ \left(\frac{1}{40} + \frac{1}{40}\right)}{154} = 1.169 = 1.081^2$$

The degrees of freedom for the error term are $abc(n - 1)$, or 72 in this design, assuming 10 subjects in each of the eight cells. The F of 1.17 is not significant on 1 and 72 *df* and therefore we lack sufficient evidence to reject the hypothesis of no interaction. In other words, we cannot conclude

that the advantage of the web format over the text format is significantly greater for experts than for novices.

The tests of the *FI* and *IE* interactions follow from the example of the test of the *FE* interaction.

The Second-Order (Three-Factor) Interaction. The eight cell means in Table 9.11 are plotted in Fig. 9.3. We assigned instructions to the x-axis, and had the experts' means in one panel and the novices in the other, with different lines for the two formats. However, this assignment of variables in the plot is arbitrary. We could have had the two formats, or the two types of instructions, in different panels. We will soon discuss some factors that may influence the decision when plotting means from a three-factor experiment. For now, let's focus on the interpretation of the second-order interaction.

In the left-hand panel of Fig. 9.3, we have plotted the interaction of format and instructions at the novice level of experience; we designate this interaction by *FI/N*. In the $2 \times 2 \times 2$ design, it is helpful to think of the three-factor interaction as a contrast of two-factor interactions. For example, in the novice panel on the left, the advantage of the web over the text format is larger in the argument than in the summary condition. However, the opposite is true in the expert panel on the right; there, the advantage of the web format over the text format is larger under summary than under argument instructions. Looking at the actual means, the *FI* interaction in the novice condition is

$$I_{FI/N} = (90.00 - 73.75) - (76.50 - 71.25)$$

The corresponding *FI* interaction in the expert condition is

$$I_{FI/E} = (89.75 - 74.00) - (88.25 - 70.50)$$

The *FIE* interaction is the difference between the two interactions; that is,

$$I_{FIE} = [(90.00 - 73.75) - (76.50 - 71.25)] - [(89.75 - 74.00) - (88.25 - 70.50)]$$

We can rewrite this as

$$I_{FIE} = (90.00 + 71.25 + 74.00 + 88.25) - (73.75 + 76.50 + 89.75 + 70.50)$$

The average of the four terms to the left of the minus sign is 80.875 and the average of the four

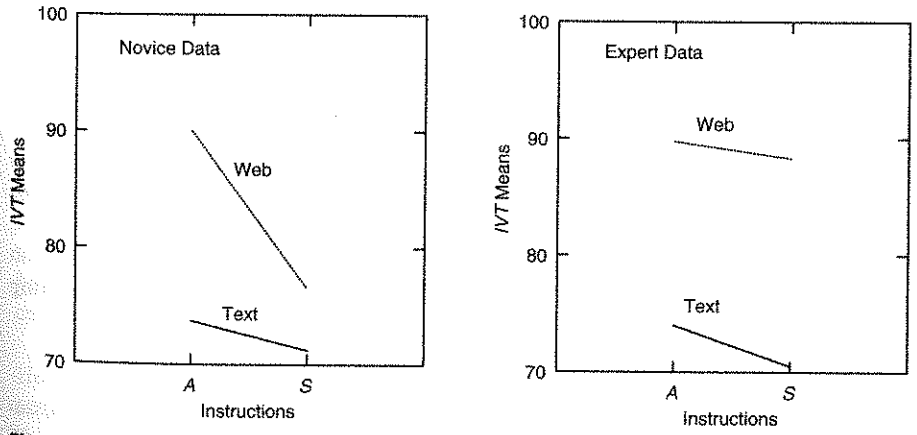

Fig. 9.3. A plot of the means in Table 9.8.

terms to the right of the minus sign is 77.625. We again have a basis for converting an interaction into a comparison of two means and therefore can test the three-way interaction with either a t test or an F test.

$$t = \frac{80.875 - 77.675}{\sqrt{154\left(\frac{1}{40} + \frac{1}{40}\right)}} = 1.171$$

or

$$F = \frac{(80.875 - 77.675)^2 \Big/ \left(\frac{1}{40} + \frac{1}{40}\right)}{154} = 1.372 = t^2$$

These significance tests are both relevant to the null hypothesis that the magnitude of the *FE* interaction does not differ for experts and novices. Based on 1 and 72 *df*, the F test does not approach significance. We conclude that the three-factor interaction is not significant. We cannot conclude that the *FI* population interaction differs as a function of the level of experience with computers. Nor does the *FE* interaction differ as a function of instructions, nor the *IE* as a function of the format. No matter which simple two-factor interactions are contrasted, the numerator of the F ratio always is equivalent to a contrast of the same two sets of four means.

9.4.4 More on 2^3 Interactions

A significant three-factor interaction means that the simple interaction effects of any two variables vary as a function of the level of the third variable. Researchers often understand this to mean that whenever the plot of the *AB* combinations looks different at different levels of *C*, the three-factor interaction is likely to be significant. However, plots like the one in Fig. 9.3 can be misleading with respect to the three-factor interaction. The following set of means should help us understand this point.

	C_1		C_2	
	B_1	B_2	B_1	B_2
A_1	22	11	34	23
A_2	20	14	23	17

Panel *a* of Fig. 9.4 presents a plot of the eight cell means under consideration. If these were population means, would you think that there is a second-order interaction? The pattern of means looks different at C_1 than at C_2; the lines cross in the C_1 panel, but not in the C_2 panel. As a result, students usually believe that an *ABC* interaction is present. In fact, if we calculate the interaction contrast, we find it is exactly zero, so there cannot be an *ABC* interaction. For example, calculating the *AB* interaction contrast at each level of *C*, and subtracting,

$$I_{ABC} = I_{AB/C_1} - I_{AB/C_2} = [(22 - 11) - (20 - 14)] - [(34 - 23) - (25 - 19)] = 0.$$

Sometimes plotting the data in different ways is helpful. In Panel *b* of Fig. 9.4, the data from Panel *a* have been replotted. Several points are now clearer than in Panel *a*. In particular, it should

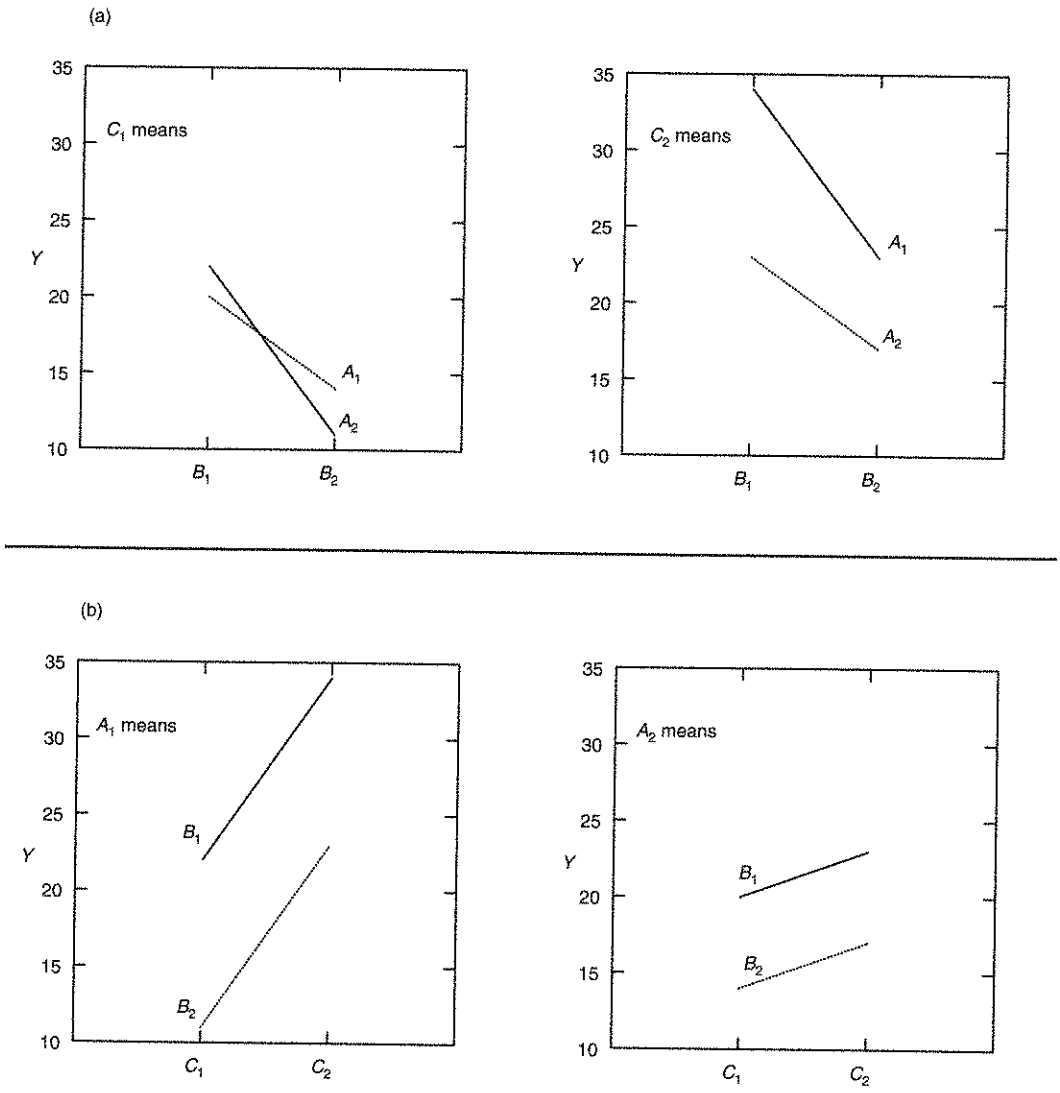

Fig. 9.4. Two ways of plotting a three-factor interaction.

be evident that there is no BC interaction, something that was not at all clear in Panel a. It also appears that there is an AB interaction because the difference between the B_1 and B_2 lines is greater in the A_1 panel than in the A_2 panel. Finally, it appears that there is no second-order interaction because the BC interaction contrast is zero in both panels. Of course, these are idealized data points, lacking the variability present in real data. However, the point still stands that it is often helpful to plot data in several ways. Different patterns may become evident, making clearer why certain effects in the ANOVA were significant whereas others were not.

The example of Panel a of Fig. 9.4 demonstrates that the pattern of means can be deceptive. However, some patterns will clearly signal the possibility of a three-factor interaction. If the lines in an AB plot are approximately parallel (i.e., there is no AB interaction) at one or more levels of C, but there is an interaction at least at one other level of C, an ABC interaction is indicated. Also, if

the lines in one panel converge whereas those in other panels diverge, an *ABC* interaction is indicated. If the two *AB* plots are the same (or displaced by a constant amount), except for one point, there is reason to expect a three-factor interaction.

9.5 MORE THAN THREE INDEPENDENT VARIABLES

The analyses of data from between-subject designs involving more than three factors are in all respects straightforward generalizations of the material presented for two- and three-factor designs. However, designs with four or more factors become unwieldy in a couple of respects. Each variable and each possible combination of variables is a potential contributor to the total variability, and so is the variability among scores within each cell of the design. As the number of factors increases, the number of possible effects to be tested increase rapidly; with f factors there are $2^f - 1$ possible effects to test. For example, with five factors there are 31 possible effects: 5 main effects, 10 two-way interactions, 10 three-way interactions, 5 four-way interactions, and 1 five-way interaction.

The calculations of higher-order interactions follow directly from what we learned for two- and three-factor interactions. As might be guessed, the degrees of freedom for any higher-order interaction are a product of the degrees of freedom for the variables entering into the interaction. For example, an *ABCD* interaction would have $(a-1)(b-1)(c-1)(d-1)$ *df*. However, although the calculations are simple, the interpretation of such higher-order interactions is often difficult. We can say that a significant four-way interaction indicates that the interaction of any three variables is a function of the level of the fourth variable, but that is not very enlightening. Unless we have prior grounds for expecting such interactions to be significant, or can attribute the interaction to some subset of cell means, care should be taken before making too much of the result.

9.6 MEASURES OF EFFECT SIZE

In Chapter 8, we introduced η^2 and ω^2 as measures of effect size. Both measures assessed the magnitude of an effect against the total variability in the experiment. There are problems with this general approach to measuring effect size when we consider multi-factor designs. Namely, the introduction of more factors into a design will generally result in an increase in total variability. As a result, the assessment of the magnitude of the effect of some factor, *A*, will decrease as the number of factors in the design increases. In short, the use of η^2 or ω^2 as measures of importance does not allow comparisons across designs with different numbers of factors.

The solution that has been offered for this problem is to use either *partial* η^2 or *partial* ω^2 to assess effect size in a design. Partial η^2 is defined as $SS_{Effect}/(SS_{Effect} + SS_{Slcells})$. Partial ω^2 is defined similarly as $\sigma^2_{Effect} / (\sigma^2_{Effect} + \sigma^2_e)$. These measures do not depend on the number of factors in a design because they explicitly exclude other factors from consideration. However, partial η^2 and partial ω^2 do not solve the problem of comparability of measures across designs.

In order to create effect size measures that are comparable across designs, we must take into account the nature of the factors in the design. We will follow the lead of Olejnik and Algina (2003), who proposed a distinction between extrinsic and intrinsic factors in a design (also, see Cohen, 1973). An *extrinsic factor* is a variable that is manipulable and therefore independent of subject characteristics and other factors in a design. For example, the dosage level of a drug or the number of presentations of an item in a memory experiment are examples of extrinsic factors. An *intrinsic factor* is a variable that cannot be manipulated, although it might be controlled (e.g., by blocking on

the factor) or measured. For example, in a learning experiment, intrinsic factors might include ability or prior experience with the task. The total variance in an experiment varies as a function of the number of extrinsic factors in the design. In contrast, total variance is not influenced by the number of intrinsic factors in a design because those factors contribute variability whether or not they are measured or controlled.

The distinction between extrinsic and intrinsic factors implies that the effect size statistic for a factor A should meet two criteria:

1. The statistic should not be affected by the contribution of extrinsic factors.
2. The baseline for assessing an effect size should include both random variance and variability associated with intrinsic factors.

Let us see how these criteria can be applied to the two measures of effect size we have been considering.

9.6.1 Eta-Squared (η^2) for the Multi-Factor Between-Subjects Design

The first criterion is readily met. For example, in a three-factor experiment in which all the factors have been manipulated, the statistic would be

$$\eta_p^2\,(Effect) = \frac{SS_{Effect}}{SS_{Effect} + SS_{Slcells}} \tag{9.11}$$

As we have seen, this statistic is referred to as *partial* η^2. We use the subscript "p" in the notation to distinguish partial η^2 from the classical η^2 defined in Chapter 8. η_p^2 is an appropriate measure of effect size in a design that contains only extrinsic factors. It is not appropriate, however, if the design contains intrinsic factors, as in the next example.

Assume that one of the three factors in the experiment is the subject's experience with the task. Factors such as experience, or gender, or level of ability are intrinsic to the subject so they should be retained in the denominator of our calculation of η^2. Olejnik and Algina (2003) propose a statistic they call *general eta-squared*, which we will notate with a "g" in the subscript to distinguish it from classical and partial eta-squared. In a design with two extrinsic factors (A, B) and one intrinsic, C, η_g^2 for A now would be

$$\eta_g^2\,(A) = \frac{SS_A}{SS_A + SS_C + SS_{AC} + SS_{BC} + SS_{ABC} + SS_{Slcells}} \tag{9.12a}$$

which is equivalent to

$$\eta_g^2\,(A) = \frac{SS_A}{SS_{total} - SS_B - SS_{AB}} \tag{9.12b}$$

The reasoning behind this equation is that if we were comparing this statistic with one based on a one-factor design, the error variance in the one-factor design would include variability due to the intrinsic factor, C, and its interaction with other factors. Therefore, in order to have comparable values of η^2 for the two designs, we include the intrinsic sources of variability in the denominator of the statistic for the two-factor design.

Now suppose we wanted a measure of the effect size for the intrinsic factor, C, in a three-factor design. Then,

$$\eta_g^2(C) = \frac{SS_C}{SS_C + SS_{AC} + SS_{BC} + SS_{ABC} + SS_{Slcells}} \tag{9.13a}$$

which can be rewritten as

$$\eta_g^2(C) = \frac{SS_C}{SS_{total} - SS_A - SS_B - SS_{AB}} \tag{9.13b}$$

We may summarize developments as follows:

The denominator of η_g^2 includes the sums of squares for the effect of interest, for the within cell error term, and for all intrinsic effects and their interactions.

This rule holds whether the effect of interest is a main or interaction effect. As one further example, consider the design of Table 9.11 in which *Format* and *Instructions* are extrinsic factors and *Experience* is an intrinsic factor. If we wanted eta-squared for the *Format* × *Instruction* interaction,

$$\eta_g^2(FI) = \frac{SS_{FI}}{SS_{FI} + SS_E + SS_{FE} + SS_{IE} + SS_{FIE} + SS_{SIFEI}} = \frac{SS_{FI}}{SS_{total} - SS_I - SS_F}$$

Using the results in Table 9.12, $\eta_g^2(FI) = 101.25/(16{,}215.5 - 551.25 - 151.25) = .007$. Clearly, the interaction makes only a small contribution. Suppose that instead of experience, our third factor had been some manipulated variable—perhaps the time allowed for studying the material. In that case, the general eta-squared formula reduces to partial eta-squared:

$$\eta_g^2(FI) = SS_{FI} / (SS_{FI} + SS_{Slcells}) = \eta_p^2(FI)$$

and the proportion of variance when all factors are extrinsic is now $101.25/(101.25 + 11088)$, or .009, slightly larger than before, but still small.

We conclude our discussion of eta-squared with a caution. Statistical packages do not compute values of general η^2; for example, SPSS outputs values of partial eta-squared. As we have argued, partial eta-squared values generally are not comparable across different experimental designs, whereas general eta-squared values are comparable. Although it is tempting to report the partial η^2 values readily available in a computer output, investigators should calculate and report values of general eta-squared.

9.6.2 Omega-Squared (ω^2) for the Multi-Factor Between-Subjects Design

Although η_g^2 is easily calculated and interpreted, it is an overestimate of the population variance attributable to a factor. As we discussed in Chapter 8, $\hat{\omega}^2$ is more satisfactory in that respect. Most commonly, *partial* ω^2 has been reported. In a multi-factor design in which we wish to measure the effect of factor A, this parameter would be defined as

$$\omega_p^2(A) = \sigma_A^2 / (\sigma_A^2 + \sigma_e^2)$$

However, we prefer to again follow the approach suggested by Olejnik and Algina (2003). Therefore, we will focus our discussion on *general* ω^2. Because our estimate of ω_g^2 requires estimates of population variances., we first define a general formula for the estimate of the variance of any main or

interaction effects, assuming that all levels of all variables have been arbitrarily selected; that is, that these are *fixed-effect variables*.[4] For such between-subjects design, the general formula is

$$\hat{\sigma}^2_{Effect} = df_{Effect} \times (MS_{Effect} - MS_{S/cells})/N \tag{9.14}$$

where N is the total number of scores. For example, for a three-factor between-subjects design, the estimate of the variance of the AB population interaction effects is

$$\hat{\sigma}^2_{AB} = (a-1)(b-1)(MS_{AB} - MS_{S/ABC})/abcn$$

Using the example of Table 9.11, the estimates of the population variances for the sources listed in Table 9.12 are

$\hat{\sigma}^2_F$	$\hat{\sigma}^2_I$	$\hat{\sigma}^2_E$	$\hat{\sigma}^2_{FI}$	$\hat{\sigma}^2_{FE}$	$\hat{\sigma}^2_{IE}$	$\hat{\sigma}^2_{FIE}$
45.341	4.966	0	0	.325	0	.716

Estimates have been set to zero when calculations based on Equation 9.14 had negative results, and σ^2_e is estimated by $MS_{S/cells}$.

Once the estimates of population variances have been calculated, we can calculate $\hat{\omega}^2$. We will follow the notational conventions introduced for η^2 to distinguish three variants of ω^2; namely, ω^2 without a subscript denotes the classical ω^2 introduced in Chapter 8, ω^2_p denotes partial ω^2, and ω^2_g denotes general ω^2. We need only a slight revision of the rule we formulated for calculating an estimate of general ω^2, $\hat{\omega}^2_g$:

The denominator of $\hat{\omega}^2_g$ includes the estimates of variances for the effect of interest, for the within cell error term, and for all intrinsic effects and their interactions.

Some examples should clarify this rule. Assume that we have three extrinsic factors; say, format, instructions, and time exposed to the material. Then, the estimate of *general* $\hat{\omega}^2$ for format is

$$\hat{\omega}^2_g(F) = \frac{\hat{\sigma}^2_F}{\hat{\sigma}^2_F + \hat{\sigma}^2_e} = \frac{(f-1)(MS_F - MS_{S/cells})/N}{(f-1)(MS_F - MS_{S/cells})/N + MS_{S/cells}}$$

Substituting either the previously calculated estimates of the variances or values of mean squares from Table 9.12, $\hat{\omega}^2_g(F) = .227$.

Now assume the design of Table 9.11 in which one factor, experience, is intrinsic. In that case,

$$\hat{\omega}^2_g(F) = \frac{\hat{\sigma}^2_F}{\hat{\sigma}^2_F + \hat{\sigma}^2_E + \hat{\sigma}^2_{FE} + \hat{\sigma}^2_{IE} + \hat{\sigma}^2_{FIE} + \hat{\sigma}^2_{error}}$$

Now the estimated variance due to experience and all its interactions contributes to the denominator. Substituting numerical estimates,

$$\hat{\omega}^2_g(F) = 45.341/(45.341 + 0 + .325 + 0 + .716 + 154) = .227$$

Because experience and its interactions contribute little variance in this experiment, the estimate is little changed from when we assumed that all factors were extrinsic. In either case, Cohen's suggested guidelines (see Chapter 8) indicate that format has a large effect.

To summarize developments thus far, we have presented an approach to computing measures of effect size that has as a goal comparability across designs. We endorse the approach advocated by

[4] This is independent of whether the factor is extrinsic or intrinsic.

Olejnik and Algina (2003) based on distinguishing extrinsic and intrinsic factors. We acknowledge that this approach is not yet widely adopted. However, general η^2 and general ω^2 do a better job of achieving the goal of comparability than do the more conventional statistics of partial η^2 and partial ω^2.

9.6.3 Cohen's *f* for the Multi-Factor Between-Subjects Design

As described in Chapter 8, Cohen (1988) defined *f* as

$$f = \sigma_{effect} / \sigma_{error} \tag{9.15}$$

For example, based on our previously calculated variance estimates,

$$\hat{f}_{Format} = \sqrt{45.341 / 154} = .54$$

As with $\hat{\omega}^2$, this is a large effect according to Cohen's guidelines. However, Equation 9.15 does not take into consideration whether the remaining factors are intrinsic or extrinsic. In order to increase compatibility across different designs, we might define *f* in a way consistent with the Olejnik and Algina (2003) approach to η^2 and ω^2. As with general ω^2, general *f* would involve a denominator that incorporates intrinsic effects, in addition to random error.[5] As with η_g^2 and ω_g^2, such a statistic would make sense in comparing effect sizes in two experiments, one of which involved a design in which intrinsic factors were controlled and one of which involved a design in which this was not the case.

Despite the advantage of f_g when comparing values of *f* based on data from different designs, we will limit our use to the classical formula for *f* in Equation 9.15. The reason for this is that *f* was viewed by Cohen as a parameter dictating the power of the *F* test, and is used in this way by the G*Power software cited earlier in this book. Because the denominator of *F* is based solely on the within-cell error variance, a value of *f* that includes intrinsic variation in its denominator would underestimate the power of the *F* test.

9.7 *A PRIORI* POWER CALCULATIONS

Assume that we wish to replicate the *Wiley–Voss* study. Further assume that our primary interest is whether the instructional effects vary. Cohen's $f = .262$ for instructions in Table 9.7. Suppose we want power = .8 to reject H_0 if the standardized effect is medium, as this value suggests. Fig. 9.5 shows the input to G*Power 3 and the resulting output. The total *N* of 179 translates to 22 subjects per condition, which is more than twice the number in the *Wiley–Voss* experiment.

9.8 UNEQUAL CELL FREQUENCIES

When cell frequencies are not equal, the ANOVA presented so far in this chapter has a problem. In this section, we describe the nature of the problem, and briefly introduce the different analyses that are available. Chapter 24 provides more detailed coverage within a regression framework.

[5] Let γ be the sum of all variance estimates involving an intrinsic factor; then, general *f* is $\hat{f}_g(Effect) = \sqrt{\hat{\sigma}^2_{Effect} / (\hat{\sigma}^2_e + \hat{\gamma})}$.

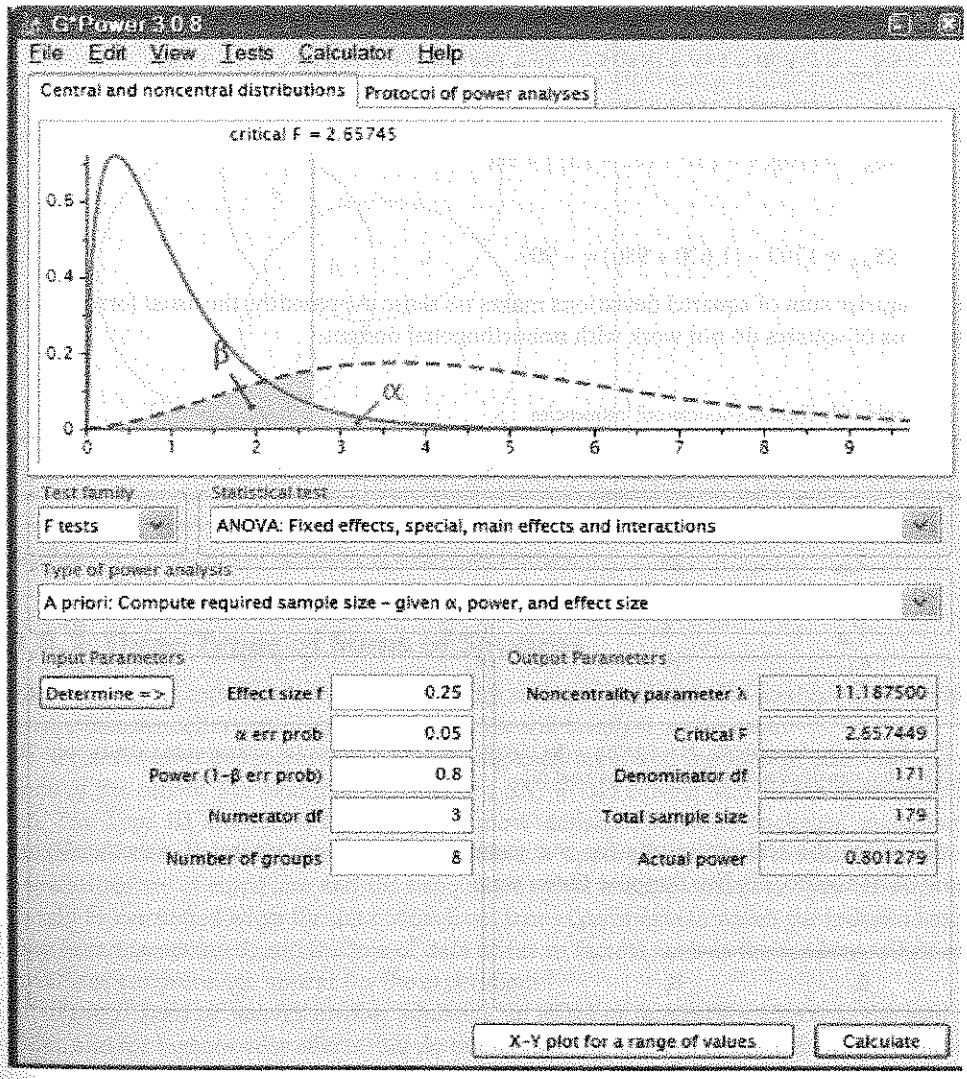

Fig. 9.5 G*Power 3 screen for *a priori* power in the multi-factor between-subjects design.

9.8.1 The Problem

In the developments thus far in this chapter, we considered only cases in which the number of scores is the same in each of the *ab* cells. As we stated in Chapter 6, when *n*s are not equal, heterogeneity of variance has a greater impact on Type 1 and 2 error rates. Furthermore, the sums of squares for main and interaction sources usually will not add to the SS_{cells} when each is calculated by ignoring the other effects. This is because when cell sizes are unequal and disproportional, the sums of squares are not independently distributed; the design is said to be *nonorthogonal*.

We can illustrate the problem by calculating sums of squares using the means in Table 9.13. Summing the squared deviations of the means about the grand mean and multiplying by the

number of scores on which each mean is based, we have

$$SS_{cells} = (2)(20 - 15)^2 + (8)(25 - 15)^2 + (8)(5 - 15)^2 + (2)(10 - 15) = 1700$$

$$SS_A = (10)[(24 - 15)^2 + (6 - 15)^2] = 1{,}620$$

$$SS_B = (10)[(8 - 15)^2 + (22 - 15)^2] = 980$$

and

$$SS_{AB} = 1700 - (1{,}620 + 980) = -900$$

Of course, a negative sum of squared deviations makes no sense. Apparently, the usual formulas for calculating sums of squares do not work with nonorthogonal designs.

Table 9.13 Example with disproportionate cell frequencies

		B_1	B_2	$n_{i.}$	$\bar{Y}_{.j.}$
A_1	n_{1k}	2	8	10	
	$\bar{Y}_{.1k}$	20	25		24
A_2	n_{2k}	8	2	10	
	$\bar{Y}_{.2k}$	5	10		6
	$n_{.k}$	10	10	$n_{..} = 20$	
	$\bar{Y}_{..k}$	8	22		$\bar{Y}_{...} = 15$

Note: $\bar{Y}_{.j.} = \Sigma_k n_{jk} \bar{Y}_{.jk} / n_{j.}$; for example, $24 = [(2)(20) + (8)(25)]/10$. The column means are computed in a similar way. The grand mean (15) is the sum of all scores divided by the total N.

The reason for the strange results for our example will become clearer if we consider an extreme case of nonorthogonality. Suppose the ns were

	B_1	B_2
A_1	0	8
A_2	8	0

Now SS_A and SS_B are identical; both are based solely on the difference between the A_1B_2 and A_2B_1 cell means, and therefore the A and B main effects are perfectly correlated. In Table 9.13, the correlation is not perfect but it is still high. The magnitude of both SS_A and SS_B will still depend primarily (though not entirely) on the difference between the A_1B_2 and A_2B_1 means.

Fig. 9.6 contains a graphic representation of the situation when cell frequencies are unequal. The square represents the SS_{total}. The circles represent SS_A, SS_B, and SS_{AB}; overlap of circles represents the covariance of effects that results from nonorthogonality. Note that when cell frequencies are equal, the three circles do not overlap. Covariances can be positive or negative so that subtracting the covariance from a sum of squares might result in a smaller quantity (if the covariance is positive) or a larger one (if the covariance is negative). The presence of correlations among effects poses choices in data analysis and interpretation.

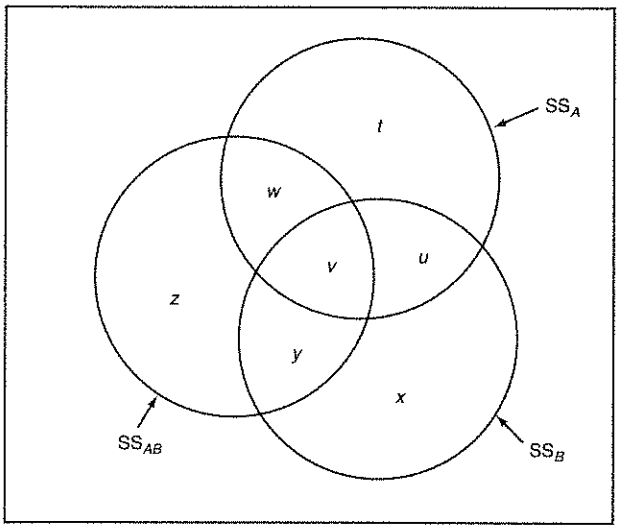

Fig. 9.6 The partitioning of variability in a two-factor design.

9.8.2 Three Types of Sums of Squares

There are three approaches available in many software packages for analyzing data when effects covary because cell frequencies are unequal. We will briefly describe each.

Type III sums of squares involve adjusting each main and interaction effect for the contributions of the others. For example, the adjusted SS_A would consist only of the area labeled t in Fig. 9.6. The Type III analysis is the default in most statistical software packages, including SPSS. It weights all the cell means equally and therefore is appropriate if it is assumed that the sampled populations are of equal size and the unequal ns reflect chance variation. Type III sums of squares should be calculated when the data come from true experiments in which the independent variables have been manipulated, and the loss of data is due to factors such as the random failure of subjects in various conditions to appear for the experiment.

Type II sums of squares involve adjusting an effect of interest for effects at the same or lower order, and for higher-order effects that do not include the effect of interest. In the two-factor design, this requires adjusting the sum of squares for each main effect for variability due to the other main effect, but not for the interaction. In Fig. 9.6, the Type II SS_A would be represented by the areas labeled t and w. This analysis implies that there are no interaction effects. Therefore, the interaction mean square can be averaged ("pooled"; see the next section) with the within-cell term, yielding more error degrees of freedom and potentially more power. However, such pooling runs the risk of failing to detect interaction effects that may exist, and of inflating the error term and thus increasing the rate of Type II errors when testing main effects. As a general rule, Type II sums of squares should not be calculated unless there is strong *a priori* reason to assume no interaction effects, and a clearly nonsignificant interaction sum of squares.

Type I sums of squares involve a hierarchical analysis. For example, suppose we are interested in the effects of educational level upon income. We might wish to control for other factors such as gender. In that case, we would first remove the sum of squares due to gender, and then calculate the sum of squares due to education. In terms of Fig. 9.6, if we represent education by A and gender by B, the sum of squares for gender would correspond to the full circle for SS_B and the education sum

of squares, adjusted for gender, would correspond to the areas t and w. Type I sums of squares rest on the assumption that cell sizes represent population sizes, an assumption that is likely to be met when the independent variable is observed rather than manipulated (e.g., income level, occupation, or clinical diagnosis) or when some treatments are more aversive than others (e.g., subjects may find some diets more difficult to maintain in a study of weight loss).

Although the Type III sum of squares is usually the default, all three types of analyses are available in many software packages. We dropped scores randomly from the eight cells of the *Wiley–Voss* data set and used SPSS to analyze the data in order to note the similarities and differences among the three types of sums of squares.

9.8.3 A Numerical Example

The data set is linked to the *Wiley–Voss* page on the book's website and has the file name W*iley Unequal_N*. The means and cell frequencies are in Table 9.14. We analyzed the data using the univariate option in the *General Linear Model* menu of SPSS, but other statistical packages provide similar options. Table 9.15 presents the results of several analyses, each after choosing a different model.

The following points should be noted about the results in Table 9.15:

1. The within-cell error term is the same in all analyses in the table. It is obtained by calculating the sums of squared deviations of scores about their cell means, and then summing across cells; the degrees of freedom are $\Sigma_j(n_j - 1) = N - a$.

Table 9.14 Cell means (and frequencies) for an experiment with unequal n

	Instructions (A)			
Format	A_1 (Narrative)	A_2 (Summary)	A_3 (Explanation)	A_4 (Argument)
B_1 (text)	68.000 (5)	75.714 (7)	70.000 (4)	73.750 (8)
B_2 (web)	76.250 (8)	75.000 (2)	76.667 (8)	90.000 (8)

Table 9.15 Sums of squares (*SS*) for data with unequal cell frequencies

SV	*df*	Type III *SS*	Type II *SS*	Type Ia *SS*	Type Ib *SS*
Instructions (A)	3	801.57 (t)	854.97 ($t + w$)	683.77 ($t + w + v + u$)	854.97 ($t + w$)
Format (B)	1	567.35 (x)	1004.54 ($x + y$)	1004.54 ($x + y$)	833.33 ($x + y + v + u$)
AB	3	368.60 (z)	368.60 (z)	368.60 (z)	368.60 (z)
Ss/AB	40	5409.76	5409.76	5409.76	5409.76

Note: The Type Ia *SS* is obtained when A is entered first in the *Univariate* dialog box; the Type Ib *SS* is the result of entering B first. The letters in parentheses refer to the corresponding areas in Fig. 9.6. See the text for further explanation.

2. The interaction term is also invariant across analyses. In all four cases, it is the sum of squares adjusted for the contributions of all other effects. This is $SS_{Between\ Groups} - (SS_A + SS_B)$ where $SS_A + SS_B$ is obtained from either Type I analysis.
3. For the Type Ia *SS*, the *A* sum of squares was calculated while ignoring all other sources, and the *B* sum of squares was calculated after adjusting for (removing variability due to) *A* effects. For the Type Ib *SS*, the *B* sum of squares was calculated while ignoring all other sources, and the *A* sum of squares was calculated after adjusting for (removing variability due to) *B* effects.
4. The Type Ia sum of squares for *B* equals the Type II sum of squares for *B*. This is because both analyses adjust the *B* sum of squares for variability due to *A*. Similarly, the Type Ib sum of squares for *A* equals the Type II sum of squares for *A* because both adjust the *A* variability for the contribution of *B*.

The set of possible analyses was performed to illustrate the differences in results. However, in analyzing data from a study, only one of these analyses should be performed. Which one will depend upon whether variability in cell frequencies is assumed to be due to chance, in which case Type III is appropriate; whether there is strong reason to believe that the interaction effects are negligible, in which case Type II is appropriate; or whether it is assumed that there is a causal relationship among the factors in the study, in which case a Type I analysis is appropriate.

9.9 POOLING IN FACTORIAL DESIGNS

Pooling is the process by which two or more mean squares are averaged. This is often done when the investigator believes that some source of variance contributes only error variance and therefore pools that mean square with the error mean square. It is also done unintentionally when the investigator fails to consider some factor in the design. Because pooling affects tests of sources of interest, it merits a closer look.

9.9.1 What Is Pooling?

When two or more sources of variance are pooled, the sums of squares are added together and divided by the sum of the degrees of freedom. For example, in a two-factor design, the pool of the *AB* term and the *S/AB* term is

$$MS_{pool} = \frac{SS_{AB} + SS_{S/AB}}{df_{AB} + df_{S/AB}} \tag{9.16}$$

Equation 9.16 can be rewritten as the weighted average of the two mean squares:

$$MS_{pool} = \left(\frac{df_{AB}}{df_{AB} + df_{S/AB}}\right) MS_{AB} + \left(\frac{df_{S/AB}}{df_{AB} + df_{S/AB}}\right) MS_{S/AB} \tag{9.17}$$

This form of the equation raises the question: When is it proper to average two mean squares? The answer is that if the two mean squares estimate the same population variance, or variances, pooling is proper. In the example of Equation 9.17, the assumption is that both MS_{AB} and $MS_{S/AB}$ estimate σ_ε^2 (i.e., $\sigma_{AB}^2 = 0$). The advantage of pooling two or more estimates of the population error variance is that MS_{pool} is distributed on more *df* than $MS_{S/AB}$; therefore, *F* tests based on the pooled error

term may be more powerful than tests based on $MS_{S/AB}$. Of course, we never *know* that $\sigma^2_{AB} = 0$. If it is not, we may lose power in using the pooled error term to test a false null hypothesis about a main effect. To see why, look again at Equation 9.17. If, contrary to the assumption upon which pooling is based, $E(MS_{AB}) = \sigma^2_e + n\theta^2_{AB}$, then the weighted average of this expectation and of σ^2_e will be larger than σ^2_e. As a result, the F test of a main effect will be *negatively biased*; the expectation of the error term involves more than just σ^2_e and there will be too many Type 2 errors. Perhaps surprisingly, when the null hypothesis about the main effect is true, there may actually be an increase in Type 1 errors. The reason for this is that if MS_{AB} and $MS_{S/AB}$ do not both estimate σ^2_e, the ratio of MS_A (or MS_B) to MS_{pool} will not be distributed as F, and the tail area may be larger than the nominal α. In view of these considerations, it is not clear when, if ever, to pool. We consider this issue next.

9.9.2 When (If Ever) to Pool

One possible approach is to apply a sometimes-pool rule; interaction terms are tested against the within-cell error term, and the two mean squares are pooled if p is greater than some criterion value. With respect to the designs of the current chapter, a study by Mead, Bancroft, and Han (1975) is relevant. For a design with two fixed-effect factors and equal cell frequencies, even with the criterion α for the preliminary test of the AB source set at .50, the sometimes-pool rule often resulted in a loss of power when the null hypothesis about the main effect was false, and an increase in Type I error rate when it was true. Therefore, in designs in which all factors have fixed effects, we recommend never pooling. We will consider whether the sometimes-pool rule is advisable in other designs when we discuss those designs.

9.9.3 Unintended Pooling

Researchers often pool terms without realizing they have done so. Typically, in such cases there are one or more treatment variables and then one "nuisance" variable that the researcher regards as irrelevant. For example, the position of a reward may appear equally often in all experimental conditions of a discrimination study, or each of several experimenters may run an equal number of subjects in each condition. Ignoring such factors is essentially pooling them with the within-cell error term. As we noted in discussing Mead et al.'s (1975) results, such pooling runs the risk of a loss of power if the null hypothesis is false, or an increased Type I error rate if it is true. The message is simple: Include all factors in the analysis, whether they are of interest or not.

9.10 ADVANTAGES AND DISADVANTAGES OF BETWEEN-SUBJECTS DESIGNS

We will consider further tests of hypotheses in between-subjects designs in the next two chapters. However, this is a good point at which to review the major advantages and disadvantages of between-subjects factorial designs:

1. Assuming equal cell frequencies, the analysis of the data from the between-subjects design is much simpler than for most other designs.
2. For any given number of scores, the error degrees of freedom in the analysis of the between-subjects design will be larger than for any comparable design.
3. The requirements of the underlying model are more easily met by between-subjects designs

than by other designs, and violations of assumptions are less likely to affect the distribution of the F ratio.

4. However, there is one major disadvantage in using between-subjects designs. Because the error variance is largely due to individual differences, designs that permit the adjustment of the error variance for the differences among individuals often will enable more powerful tests and more precise estimates of population parameters, and often will do so with fewer data.

Because of the large error variance associated with between-subjects designs, these designs are most useful whenever subjects are relatively similar with respect to the dependent variable, or whenever a large N is available to compensate somewhat for the variability among individuals. Also, there are many experiments in which the nature of the independent variable (e.g., the method of instruction, or educational level of subjects) will constrain the design. However, there are many situations in which several different designs are feasible and desirable. In Chapter 13, we will introduce the idea of design efficiency, and consider the efficiency of alternative designs, and in Chapters 14–17, we will consider those designs and their analyses more closely.

9.11 SUMMARY

In this chapter, we considered experiments in which each subject contributed one score to a cell in a design in which all possible combinations of two or more factors were represented. Within this context, we covered the following topics:

- *The extension of the structural model to multi-factor designs.* Building on that model, we partitioned the total variability and degrees of freedom into components representing main, interaction, and error sources of variance.
- *The interpretation of interaction.* An interaction occurs when the magnitude of an effect varies over levels of another variable. One of the advantages of a multi-factor design is the ability to examine interactions among variables.
- *Measures of effect size.* We extended measures of importance to multi-factor designs, introducing modifications of the eta-squared and omega-squared measures presented in Chapter 8. In addition, we again considered Cohen's f.
- *Analyses when cell frequencies are unequal.* We explained why this may cause problems for the standard ANOVA, and briefly described alternative ways of calculating sums of squares, depending on what is assumed about population sizes and the presence of interaction, as well as the purpose of the study.

Thus far, we have focused on the omnibus F test. However, in most instances, other comparisons of means are of more interest. We turn next to that topic.

EXERCISES

9.1 (a) Perform an analysis of variance on the following data set (also available from the Exercises page on the book's website in the file *EX9_1*).