Chapter 17—Factorial Analysis of Variance

17.1 Thomas and Wang (1996) study:

a) This design can be characterized as a 3×2 factorial, with 3 levels of Strategy and 2 levels of delay.
b) I would expect that recall would be better when subjects generated their own key words, and worse when subjects were in the rote learning condition. I would also expect better recall for the shorter retention interval. (But what do I know?)
c)

```
Summaries of     RECALL
By levels of     STRATEGY
                 DELAY
Variable        Value  Label              Mean    Std Dev    Cases

For Entire Population                 11.602564   7.843170     78

STRATEGY       1.0000                  9.461538   6.906407     26
  DELAY        1.0000                 14.923077   5.330127     13
  DELAY        2.0000                  4.000000   2.516611     13

STRATEGY       2.0000                 11.269231   9.606488     26
  DELAY        1.0000                 20.538462   1.983910     13
  DELAY        2.0000                  2.000000   1.471960     13

STRATEGY       3.0000                 14.076923   6.183352     26
  DELAY        1.0000                 15.384615   5.454944     13
  DELAY        2.0000                 12.769231   6.796492     13
```

17.3 Analysis of variance on data in Exercise 17.1:

```
RECALL      by   STRATEGY
                 DELAY

                 UNIQUE sums of squares
                 All effects entered simultaneously
                           Sum of                Mean              Sig
Source of Variation        Squares     DF       Square        F    of F

Main Effects              2510.603      3      836.868    42.992   .000
   STRATEGY                281.256      2      140.628     7.224   .001
   DELAY                  2229.346      1     2229.346   114.526   .000

2-Way Interactions         824.538      2      412.269    21.179   .000
   STRATEGY DELAY          824.538      2      412.269    21.179   .000

Explained                 3335.141      5      667.028    34.267   .000

Residual                  1401.538     72       19.466

Total                     4736.679     77       61.515
```

There are significant differences due to both Strategy and Delay, but, more importantly, there is a significant interaction.

> This is a good example for showing all three effects. The Delay and Interaction effects are obvious, but the overall Strategy effect is harder to see. You would do well to calculate the Strategy means, which are 9.46, 11.27, and 14.08, respectively. It will help if you draw those means on the figure for Exercise 17.2.

.

17.5 Bonferroni tests to clarify simple effects for data in Exercise 17.4:

$$t = \frac{\overline{X}_i - \overline{X}_j}{\sqrt{\frac{MS_{error}}{n_i} + \frac{MS_{error}}{n_j}}}$$

For Data at 5 Minutes Delay:

For Generated versus Provided:

$$t = \frac{14.92 - 20.54}{\sqrt{\frac{20.7009}{13} + \frac{20.7009}{13}}} = \frac{-5.62}{1.784} = -3.15$$

For Generated versus Rote:

$$t = \frac{14.92 - 15.38}{\sqrt{\frac{20.7009}{13} + \frac{20.7009}{13}}} = \frac{-0.46}{1.784} = -0.26$$

For Provided versus Rote:

$$t = \frac{20.54 - 15.38}{\sqrt{\frac{18.2308}{13} + \frac{18.2308}{13}}} = \frac{5.16}{1.784} = 2.89$$

For Data at 2 Day Delay:

For Generated versus Provided:

$$t = \frac{4.00 - 2.00}{\sqrt{\frac{18.2308}{13} + \frac{18.2308}{13}}} = \frac{2.00}{1.674} = 1.19$$

For Generated versus Rote:

$$t = \frac{4 - 12.77}{\sqrt{\frac{18.2308}{13} + \frac{18.2308}{13}}} = \frac{-8.77}{1.674} = -5.24$$

For Provided versus Rote:

$$t = \frac{2 - 12.77}{\sqrt{\frac{20.7009}{13} + \frac{20.7009}{13}}} = \frac{-10.77}{1.674} = -6.43$$

For 6 comparisons with 36 *df*, the critical value of *t* is 2.80.

For the 5-minute delay, the condition with the key words provided by the experimenter is significantly better than both the condition in which the subjects generate their own key words and the rote learning condition. The latter two are not different from each other.

For the 2-day delay, the rote learning condition is better than either of the other two conditions, which do not differ between themselves.

We clearly see a different pattern of differences at the two delay conditions. The most surprising result (to me) in the superiority of rote learning with a 2 day interval.

> In running these Bonferroni tests, I had a choice. I could have thought of each simple effect as a family of comparisons, and obtained the critical value of *t* with 3 comparisons for each. Instead I chose to treat the whole set of 6 comparisons as a family and adjust the Bonferroni for 6 tests. There is no hard and fast rule here, and many might do it the other way. The results would not change regardless of what I decided.

17.7 The results in the last few exercises have suggested to me that if I were studying for a Spanish exam, I would fall back on rote learning, painful as it sounds and as much against common wisdom as it is.

17.9 In this experiment we have as many primiparous mothers as multiparous ones, which certainly does not reflect the population. Similarly, we have as many LBW infants as full-term ones, which is certainly not a reflection of reality. The mean for primiparous mothers is based on an equal number of LBW and full-term infants, which we know is not representative of the population of all primiparous births. Comparisons between groups are still legitimate, but it makes no sense to take the mean of all primiparous moms combined as a reflection of any meaningful population mean.

> Many of our experiments are run this way (with equal sample sizes across groups that are not equally represented in the population), and it is important to distinguish between the legitimacy of between group comparisons and the legitimacy of combined means.

17.11 Simple effects versus *t* tests for Exercise 17.10.

a) If I had run a *t* test between those means my result would simply be the square root of the $F = 1.328$ that I obtained.

b) If I used MS_{error} for my estimated error term it would give me a *t* that is the square root of the *F* that I would have had if I had used the overall MS_{error}, instead of the MS_{error} obtained in computing the simple effect.

17.13 Analysis of variance for Spilich *et al*. Study:

Tests of Between-Subjects Effects

Dependent Variable: ERRORS

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	31744.726[a]	8	3968.091	36.798	.000
Intercept	45009.074	1	45009.074	417.389	.000
TASK	28661.526	2	14330.763	132.895	.000
SMKGRP	354.548	2	177.274	1.644	.197
TASK * SMKGRP	2728.652	4	682.163	6.326	.000
Error	13587.200	126	107.835		
Total	90341.000	135			
Corrected Total	45331.926	134			

a. R Squared = .700 (Adjusted R Squared = .681)

The main effect of Task and the interaction are significant. The main effect of Task is of no interest because there is no reason why different tasks should be equally difficult., We don't care about the main effect of Smoking either because it is created by large effects for two levels of Task and no effect for the third. What is important is the interaction.

This is a good example of a situation in which main effects are of little interest. For example, saying that smoking harms performance is not really accurate. Smoking harms performance on some tasks, but not on others. Often main effects are still interpretable in the presence of an interaction, but not here.

17.15 Simple effects to clarify the Spilich *et al*. Example.

We have already seen these simple effects in Chapter 16, in Exercises 16.18, 16.19, and 16.21.

17.17 Factorial analysis of the data in Exercise 16.2:

Tests of Between-Subjects Effects

Dependent Variable: SCORE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1059.800[a]	3	353.267	53.301	.000
Intercept	5017.600	1	5017.600	757.056	.000
AGE	115.600	1	115.600	17.442	.000
HILO	792.100	1	792.100	119.512	.000
AGE * HILO	152.100	1	152.100	22.949	.000
Error	238.600	36	6.628		
Total	6316.000	40			
Corrected Total	1298.400	39			

a. R Squared = .816 (Adjusted R Squared = .801)

Here we see that we have a significant effect due to age, with younger subjects outperforming older subjects, and a significant effect due to the level of processing, with better recall of material processed at a higher level. Most importantly, we have a significant interaction, reflecting the fact that there is no important difference between younger and older subjects for the task with low levels of processing, but there is a big difference when the task calls for a high level of processing—younger subjects seem to benefit more from that processing (or do more of it).

17.19 Nurcombe et al study of maternal adaptation.

Source	df	SS	MS	F
E (Education)	1	67.69	67.69	6.39*
G (Group)	2	122.79	61.40	5.80*
EG	2	20.38	10.19	<1
Error	42	444.62	10.59	
Total	47	655.48		

*$p < .05$

b) The program worked as intended and there was no interaction between groups and educational level.

17.21 Effect size for Level of Processing in Exercise 17.17

$$\hat{d} = \frac{\bar{X}_{Hi} - \bar{X}_{low}}{\sqrt{MS_{error}}} = \frac{15.65 - 6.75}{\sqrt{6.628}} = \frac{8.90}{2.5739} = 3.46$$

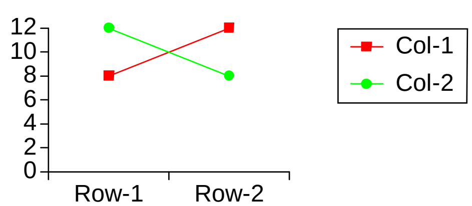
This is a very large effect size, but the data show an extreme difference between the two levels of processing.

I used the square root of MS_{error} here because that was in line with what I did in the text. But a good case could be made for adding Age and the interaction sums of squares back in and calculating a new error term. That would produce

$$\hat{d} = \frac{\bar{X}_{Hi} - \bar{X}_{Low}}{\sqrt{MS_{error-revised}}} = \frac{15.65 - 8.90}{\sqrt{13.323}} = \frac{8.90}{3.65} = 2.44$$

which is considerably smaller but still a very large effect.

17.23 Set of data for a 2×2 design with no main effects but an interaction:



Cell means:

8	12
12	8

17.25 Magnitude of effect for Exercise 17.1

Summary table from Exercise 17.1:

Source	***df***	***SS***	***MS***	***F***
Strategy	2	281.256	140.628	7.224
Delay	1	2229.346	2229.346	114.526
S × D	2	824.538	412.269	21.179
Error	72	1401.538	19.466	
Total	77	4736.679		

$$\eta^2_{Strategy} = \frac{SS_{Strategy}}{SS_{total}} = \frac{281.256}{4736.679} = .06$$

$$\omega^2_{Strategy} = \frac{SS_{Strategy} - (s-1)MS_{error}}{SS_{total} + MS_{error}} = \frac{281.256 - (3-1)19.466}{4736.679 + 19.466} = .05$$

$$\eta^2_{Delay} = \frac{SS_{Delay}}{SS_{total}} = \frac{2229.346}{4736.679} = .47$$

$$\omega^2_{Delay} = \frac{SS_{Delay} - (d-1)MS_{error}}{SS_{total} + MS_{error}} = \frac{2229.346 - (2-1)19.466}{4736.679 + 19.466} = .46$$

$$\eta^2_{SD} = \frac{SS_{SD}}{SS_{total}} = \frac{824.538}{4736.679} = .17$$

$$\omega^2_{SD} = \frac{SS_{SD} - (s-1)(d-1)MS_{error}}{SS_{total} + MS_{error}} = \frac{824.538 - (3-1)(2-1)19.466}{4736.679 + 19.466} = .16$$

17.27 Magnitude of effect for Exercise 17.13:

Summary table from Exercise 17.13:

Source	***df***	***SS***	***MS***	***F***
Task	2	28661.526	14330.763	132.895
SmokeGrp	2	1813.748	906.874	8.41
T × S	4	1269.452	317.363	2.943
Error	126	13587.200	107.835	
Total	134	45331.926		

$$\eta^2_{Task} = \frac{SS_{Task}}{SS_{total}} = \frac{28661.526}{45331.926} = .63$$

$$\omega^2_{Task} = \frac{SS_{Task} - (t-1)MS_{error}}{SS_{total} + MS_{error}} = \frac{28661.526 - (3-1)107.835}{45331.926 + 107.835} = .63$$

$$\eta^2_{Smoke} = \frac{SS_{Smoke}}{SS_{total}} = \frac{1813.748}{45331.926} = .04$$

$$\omega^2_{Smoke} = \frac{SS_{Smoke} - (s-1)MS_{error}}{SS_{total} + MS_{error}} = \frac{1813.748 - (3-1)107.835}{45331.926 + 107.835} = .04$$

$$\eta^2_{TS} = \frac{SS_{TS}}{SS_{total}} = \frac{1269.452}{45331.926} = .03$$

$$\omega^2_{TS} = \frac{SS_{TS} - (t-1)(s-1)MS_{error}}{SS_{total} + MS_{error}} = \frac{1269.452 - (3-1)(3-1)107.835}{45331.926 + 107.835} = .02$$

17.29 The two magnitude of effect measures (η^2 and ω^2) will agree when the error term is small relative to the effect in question, and will disagree when there is a substantial amount of error relative to the effect. But notice that this is a comparison of MS_{error} and a sum of squares, and sums of squares can be large when there are many degrees of freedom for them. So to some extent, all other things equal, the two terms will be in closer agreement when there are several degrees of freedom for the treatment effect.

17.31 You should restrict the number of simple effects you examine to those in which you are particularly interested (on *a priori* grounds), because the familywise error rate will increase as the number of tests increases.

> Although we routinely talk about familywise error rates with respect to multiple comparison procedures, they really apply whenever you run more than one test, whether you consider them tests on main effects and interactions, or tests on simple effects, or tests on multiple contrasts. A test is a test as far as the error rate is concerned.

Source	***df***	***SS***	***MS***	***F***
Gender	1	223.49	223.49	10.78
Condition	1	1.35	1.35	<1
G × X	1	0.69	0.69	<1
Error	56	1161.44	20.74	