2xk 2-Factor Between Groups ANOVA with EMMEANS Follow-ups

The purpose of this study was to examine the relationships of exam Review Attendance and Practice Difficulty with exam performance. Practice Difficulty was a 3-condition variable - practice problems were either about the same difficulty as the exam problems (=1), they were easier than the exam problems (=2), or they were more difficult than the exam problems (=3). Different sections of the course were randomly assigned to receive the three difficulty levels. The schedule showed the class meeting during which the exam review would occur & student's attendance was recorded. The dependent variable was performance on an examination.

Process:
There are a lot of steps to a complete analysis of a 2-way design. Different patterns of significant and non-significant effects will require different subsets of these. Here's a preview…

Initial Analysis
- Get descriptive means, plots & F-tests
- Determine what effects are significant
- Consider what main effects are likely to be interesting – based on the aggregations involved

2-way Interactions
- Get 2-way cell means & follow-up analyses to describe the 2-way interaction

Main Effects
- Get estimated marginal means & follow-up analyses to describe each main effect
- Why are the "Descriptive" and "Estimated" marginal means different ?

Initial Analysis

Get descriptive means, plots & F-tests

```
unianova testperf by ar1y2n pg1e2h3s
```
← lists DV "by" IVs
order determines left-to-right ordering of IVs in the Descriptive Statistics table

```
 / method = sstype(3)
```
← corrects each effect for all other effects

```
 / print descriptives parameters
```
← get descriptive cell and marginal means

```
 / plot profile (pg1e2h3s*ar1y2n)
```
← get plot of cell means (x-axis * "separate lines")

```
 / design = pg1e2h3s ar1y2n pg1e2h3s*ar1y2n.
```
← specify the design including the interaction that is automatically calculates from the IVs specified above)

Descriptive Statistics

Dependent Variable: pestperf

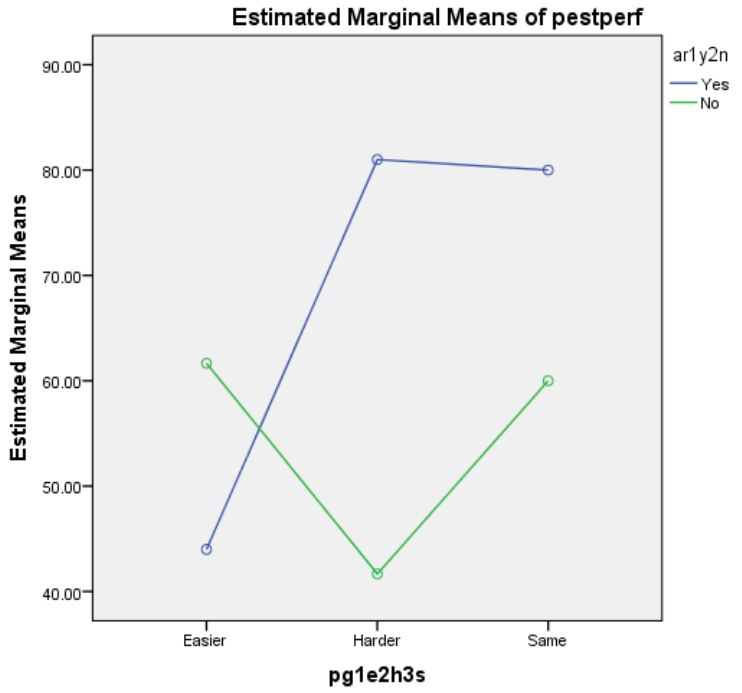
ar1y2n	pg1e2h3s	Mean	Std. Deviation	N
Yes	Easier	44.0000	9.66092	10
	Harder	81.0000	13.70320	10
	Same	80.0000	8.94427	6
	Total	66.5385	21.15510	26
No	Easier	61.6667	9.83192	6
	Harder	41.6667	11.69045	6
	Same	60.0000	8.16497	10
	Total	55.4546	12.62170	22
Total	Easier	50.6250	12.89381	16
	Harder	66.2500	23.34522	16
	Same	67.5000	12.90995	16
	Total	61.4583	18.44942	48

The "Descriptive Statistics" are the raw or "uncorrected" means.

The marginal means are weighted by the differential sizes of the cell means being aggregated.

For example, the marginal mean for the Easier PractDif is

((44.00 * 10) + (61.667 * 6)) / 16 = 50.625



From the means and the plots, it looks like Harder and Same difficulty practice work best for those who do attend the review, but Easier and Same difficulty practice work best for those who do not attend the review.

Tests of Between-Subjects Effects

Dependent Variable: testperf

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	11301.250[a]	5	2260.250	20.212	.000
Intercept	169586.806	1	169586.806	1516.532	.000
pg1e2h3s	2210.278	2	1105.139	9.883	.000
ar1y2n	2170.139	1	2170.139	19.406	.000
pg1e2h3s * ar1y2n	6301.944	2	3150.972	28.178	.000
Error	4696.667	42	111.825		
Total	197300.000	48			
Corrected Total	15997.917	47			

a. R Squared = .706 (Adjusted R Squared = .671)

We have significant effects “all around” !

Consider what lower-order effects we will need to check for descriptive/misleading patterns

Because of the significant 2-way, the means patterns of each main effect will have to be carefully checked against the corresponding simple effects to determine if they are descriptive or misleading. Remember, this will have to be done whether the main effect is significant or not – main effect nulls can be misleading!

Consider what lower-order effects are likely to be interesting – based on the aggregations involved

PractDif

- These conditions are really pretty arbitrary.
- More importantly, it is unclear what population is represented by an average of those who attended and not attend the review session!
- So, this main effect is only likely to be interesting if that main effect is descriptive, and so, it describes the behavior of both those who did and did not attend the review.

Attend the Review

- This is a straightforward operationalization of a simple variable
- However, the marginal means are of dubious value, because the PractDif conditions are arbitrary, and so it is not clear what population would be represented by the aggregate of the easier, harder, and similar difficulty performances
- So, this main effect is only likely to be interesting if that main effect is descriptive, and so, it describes the behavior of those who practiced with similarly difficult, harder, and easier materials.

***Remember – – non-significant lower-order effects that are involved in a significant higher order effect must be compared to the corresponding simple effects, to determine whether they are descriptive or misleading*!!!**

2-way Interaction

Pairwise Comparisons

You will usually want both sets of simple effects. One of those sets will be used to describe the pattern of the significant interaction. Each set will be used to determine if the corresponding main effect pattern is descriptive or misleading.

Select the set of simple effects that most directly addresses the research question or research hypothesis

The statement that, “We wanted to know if the relative difficulty of the practice material was related to test performance, and if this effect was different for those who did and did not attend the review session.” makes the selection of the simple effects to use to describe the interaction straightforward.

From this, we’ll want to focus on the simple effect of practice difficulty (easier, harder, similar) and then examine how this simple effect is different those who did and did not attend the review session.

Obtaining and describing the pairwise simple effects of Practice Difficulty for each level of Review Attendance

/ emmeans tables (ar1y2n by pg1e2h3s) compare (pg1e2h3s)

← this asks for the an analysis of the cell means for the 2-way interaction
← the order of the variables in parenthesis of the "table" command controls the display of the means
← the variable specified in the "compare" command tells which set of simple effects to test

Estimates

Dependent Variable: testperf

ar1y2n	pg1e2h3s	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Yes	Easier	44.000	3.344	37.251	50.749
	Harder	81.000	3.344	74.251	87.749
	Same	80.000	4.317	71.288	88.712
No	Easier	61.667	4.317	52.954	70.379
	Harder	41.667	4.317	32.954	50.379
	Same	60.000	3.344	53.251	66.749

Same cell means as in the Descriptives table above, but rearranged to match the tables command.

Univariate Tests

Dependent Variable: testperf

ar1y2n		Sum of Squares	df	Mean Square	F	Sig.
Yes	Contrast	8258.462	2	4129.231	36.926	.000
	Error	4696.667	42	111.825		
No	Contrast	1578.788	2	789.394	7.059	.002
	Error	4696.667	42	111.825		

Each F tests the simple effects of pg1e2h3s within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

The F-tests tell us that there is a significant simple effect of Practice Difficulty for each condition of Review Attendance.

With only 3 Practice Difficulty conditions, we will need follow-up analyses to explicate the pattern of these simple effects.

Pairwise Comparisons

Dependent Variable: testperf

ar1y2n	(I) pg1e2h3s	(J) pg1e2h3s	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
						Lower Bound	Upper Bound
Yes	Easier	Harder	-37.000*	4.729	.000	-46.544	-27.456
		Same	-36.000*	5.461	.000	-47.020	-24.980
	Harder	Easier	37.000*	4.729	.000	27.456	46.544
		Same	1.000	5.461	.856	-10.020	12.020
	Same	Easier	36.000*	5.461	.000	24.980	47.020
		Harder	-1.000	5.461	.856	-12.020	10.020
No	Easier	Harder	20.000*	6.105	.002	7.679	32.321
		Same	1.667	5.461	.762	-9.354	12.687
	Harder	Easier	-20.000*	6.105	.002	-32.321	-7.679
		Same	-18.333*	5.461	.002	-29.354	-7.313
	Same	Easier	-1.667	5.461	.762	-12.687	9.354
		Harder	18.333*	5.461	.002	7.313	29.354

Based on estimated marginal means

*. The mean difference is significant at the .050 level.

b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

The pairwise effects describing the interaction are:

	Easier v Same	Easier v Harder	Same v Harder
Did attend the review	44.0 < 80.0	44.0 < 81.0	80.0 = 81.0
Did not attend review	61.7 = 60.0	61.7 > 41.7	60.0 > 41.7

This interaction pattern allows us to anticipate that the main effect pattern of Practice Difficulty will be **misleading**

Obtaining and describing the pairwise simple effects of Review Attendance for each level of Practice Difficulty

/ emmeans tables (pg1e2h3s by ar1y2n) compare (ar1y2n)

← this asks for the an analysis of the cell means for the 2-way interaction
← the order of the variables in parenthesis of the "table" command controls the display of the means
← the variable specified in the "compare" command tells which set of simple effects to test

Estimates

Dependent Variable: testperf

pg1e2h3s	ar1y2n	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Easier	Yes	44.000	3.344	37.251	50.749
	No	61.667	4.317	52.954	70.379
Harder	Yes	81.000	3.344	74.251	87.749
	No	41.667	4.317	32.954	50.379
Same	Yes	80.000	4.317	71.288	88.712
	No	60.000	3.344	53.251	66.749

The cell means will be the same as given in the "Descriptive Statistics" above.

The F-tests tell us that the simple effect of Review Attendance is significant Same but not Easier Practice.

Univariate Tests

Dependent Variable: testperf

pg1e2h3s		Sum of Squares	df	Mean Square	F	Sig.
Easier	Contrast	1170.417	1	1170.417	10.466	.002
	Error	4696.667	42	111.825		
Harder	Contrast	5801.667	1	5801.667	51.881	.000
	Error	4696.667	42	111.825		
Same	Contrast	1500.000	1	1500.000	13.414	.001
	Error	4696.667	42	111.825		

Each F tests the simple effects of ar1y2n within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

With only 2 Review Attendance conditions, the pairwise comparisons are redundant with the F-tests.

Example.

Easier $t^2 = (-17.667 / 5.461)^2 = 10.466 = F$

Pairwise Comparisons

Dependent Variable: testperf

pg1e2h3s	(I) ar1y2n	(J) ar1y2n	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
						Lower Bound	Upper Bound
Easier	Yes	No	-17.667*	5.461	.002	-28.687	-6.646
	No	Yes	17.667*	5.461	.002	6.646	28.687
Harder	Yes	No	39.333*	5.461	.000	28.313	50.354
	No	Yes	-39.333*	5.461	.000	-50.354	-28.313
Same	Yes	No	20.000*	5.461	.001	8.980	31.020
	No	Yes	-20.000*	5.461	.001	-31.020	-8.980

Based on estimated marginal means

*. The mean difference is significant at the .050 level.

b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

The pattern of the interaction is:

Easier Practice
No Review **>** Review

Harder Practice
No Review **<** Review

Same Difficulty Practice
No Review **<** Review

This interaction pattern allows us to anticipate that the main effect of Review Attendance will be **misleading**

Alternative Analysis of Cell Means

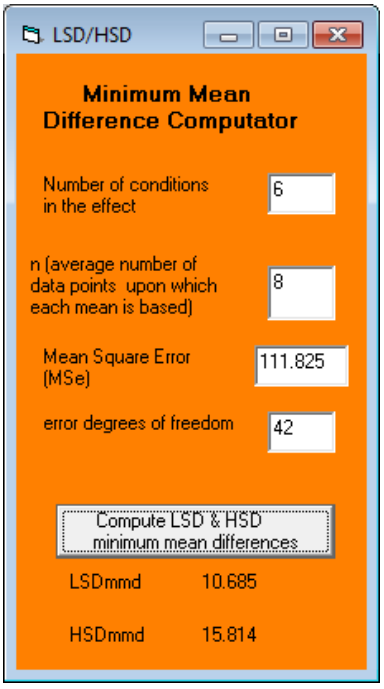
This is a BG model, so all the F-tests and follow-up analyses are based on a single error term (MSe=111.852), though the Standard Errors of the follow-ups vary with sample size. Why care? Because, the follow-up analyses are based on a t-test (that isn't shown in the output) that uses the standard error in the denominator.

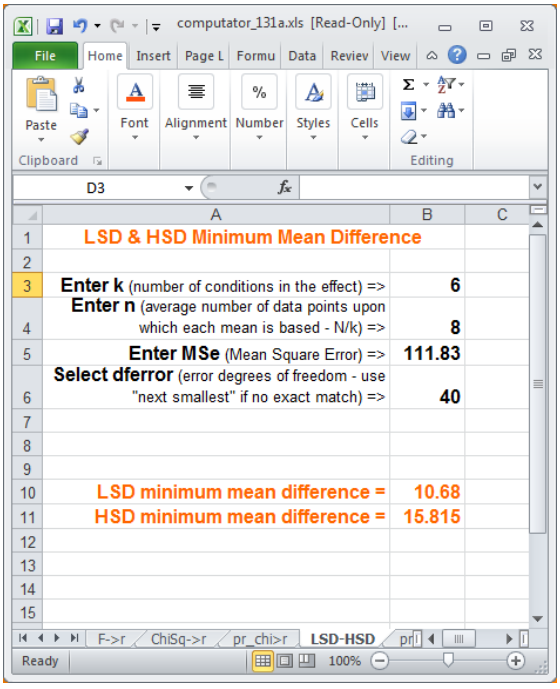
So, depending on whether the cells being compared have larger or smaller sample sizes, the standard error can be larger (smaller ns) or smaller (larger ns), and the same cell mean difference can be significant for one comparison and not significant for another!!

An alternative is to use this "full model error term" as the basis for computing an LSD value that is then used to compare any two cell means. This is an extension of the "homogeneity of variance" assumption that is made when we compute the ANOVA error term for BG models. That assumption is that it makes sense to combine the within-group variability from the different design cells, because they each represent a sample taken from different populations that all have the same variability, so the aggregate of them all is the best estimate of the variability of each. The extension in the "full model error term" approach is that since the best estimate is derived from using the full design sample, the significance test should be based on the df from all the participants.

Why do people who like this approach like it?

1. It is based on the same estimate of variability, but larger sample size, and, so, uses a smaller standard error than the pairwise error term approach. So, it provides a more powerful significance test, and more pairwise cell mean comparisons are significantly different using this approach (though the reverse can happen on occasion).
2. This approach allows the comparison of nonadjacent cells means. We might have the research hypothesis that those who attend the review and do the easier practice (mean = 44.000) have the same performance as those who do not attend the review and do the harder practice (mean = 41.667), there is no easy to get SPSS to provide this significance test, but the Computators will give us an LSDmmd that we can use to compare these means. Using the LSDmmd value, we would conclude these two groups have equivalent performances.





Another approach to testing simple effects that shows up in many examples is to use the “split file” option in SPSS and run separate analyses for each partition of the design.

```
temporary.                                  ← specify that sort & split commands will only apply to the next analysis command

SORT CASES  BY pg1e2h3s.                    ← sorts the cases by the selection variable
SPLIT FILE LAYERED BY pg1e2h3s.             ← splits the cases by the selection variable

UNIANOVA testperf  BY  ar1y2n               ← specify DV  “by”  IV (simple effect variable)
  /DESIGN = ar1y2n.
```

.

Tests of Between-Subjects Effects

Dependent Variable: testperf

pg1e2h3s	Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Easier	Corrected Model	1170.417[a]	1	1170.417	12.382	.003
	Intercept	41870.417	1	41870.417	442.962	.000
	ar1y2n	1170.417	1	1170.417	12.382	.003
	Error	1323.333	14	94.524		
	Total	43500.000	16			
	Corrected Total	2493.750	15			
Harder	Corrected Model	5801.667[b]	1	5801.667	34.223	.000
	Intercept	56426.667	1	56426.667	332.854	.000
	ar1y2n	5801.667	1	5801.667	34.223	.000
	Error	2373.333	14	169.524		
	Total	78400.000	16			
	Corrected Total	8175.000	15			
Same	Corrected Model	1500.000[c]	1	1500.000	21.000	.000
	Intercept	73500.000	1	73500.000	1029.000	.000
	ar1y2n	1500.000	1	1500.000	21.000	.000
	Error	1000.000	14	71.429		
	Total	75400.000	16			
	Corrected Total	2500.000	15			

a. R Squared = .469 (Adjusted R Squared = .431)

b. R Squared = .710 (Adjusted R Squared = .689)

c. R Squared = .600 (Adjusted R Squared = .571)

Notice that for each analysis…

The SS effect (ar1y2n) are the same.

The df-error are different

- Based on cell sample sizes for the split-analysis approach
- Based on full design N for the emmeans approach

The MSerror are different

- Based select portions of the data for the split-analysis approach
- Based on the full design for the emmeans approach

If you take the MSeffect from the split analyses and apply the MSerror from the full model, you get the same F-test as from the emmeans approach.

For the Easier split

1170.417 / 111.825 = 10.466

Here are the corresponding simple effects F-tests from the emmenans analysis,

Univariate Tests

Dependent Variable: testperf

pg1e2h3s		Sum of Squares	df	Mean Square	F	Sig.
Easier	Contrast	1170.417	1	1170.417	10.466	.002
	Error	4696.667	42	111.825		
Harder	Contrast	5801.667	1	5801.667	51.881	.000
	Error	4696.667	42	111.825		
Same	Contrast	1500.000	1	1500.000	13.414	.001
	Error	4696.667	42	111.825		

Each F tests the simple effects of ar1y2n within each level combination of the other effects shown. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

Here is the syntax to get the simple effects of practice difficulty for each review attendance condition, with LSD follow-ups (since there are 3 groups).

```
temporary.

SORT CASES  BY ar1y2n.
SPLIT FILE LAYERED BY pg1e2h3s.

UNIANOVA testperf  BY  pg1e2h3s
  /POSTHOC = pg1e2h3s (LSD)
  /DESIGN = pg1e2h3s.
```

Describing the Main Effect of Practice Difficulty

/ emmeans tables (pg1e2h3s) compare (pg1e2h3s)

Estimates

Dependent Variable: testperf

pg1e2h3s	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Easier	52.833	2.730	47.323	58.343
Harder	61.333	2.730	55.823	66.843
Same	70.000	2.730	64.490	75.510

Univariate Tests

Dependent Variable: testperf

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	2210.278	2	1105.139	9.883	.000
Error	4696.667	42	111.825		

The F tests the effect of pg1e2h3s. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

Pairwise Comparisons

Dependent Variable: testperf

(I) pg1e2h3s	(J) pg1e2h3s	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Easier	Harder	-8.500*	3.861	.033	-16.293	-.707
	Same	-17.167*	3.861	.000	-24.959	-9.374
Harder	Easier	8.500*	3.861	.033	.707	16.293
	Same	-8.667*	3.861	.030	-16.459	-.874
Same	Easier	17.167*	3.861	.000	9.374	24.959
	Harder	8.667*	3.861	.030	.874	16.459

Based on estimated marginal means

*. The mean difference is significant at the .050 level.

b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

You should notice the means shown here are not the same as the marginal means from the "Descriptive Statistics" above (50.6 for Easier, 67.5 for Same and 66.3 for Harder).

The F-test matches what's in the ANOVA table above, because both are for the corrected or unique contribution of this main effect to the model. Said differently, both are testing the mean difference among the estimated marginal means of the groups, after correcting for the other effects in the model.

The pairwise comparisons show the pattern of the main effect of Practice Difficulty to be:

Easier < Harder < Same

However, we know from the pattern of the interaction that this is not descriptive, neither for those to attended the review nor for those who did not attend the review.

This main effect must be communicated carefully, because it is potentially misleading.

Alternative Analyses of Marginal Means of Practice Difficulty

You will sometimes see folks obtain an LSDmmd value and use it to compare the marginal means, to test and describe the pattern of the main effect. That LSDmmd value will differ from the value used to compare cell means above, because the n for the marginal means is different from the n of the cell means.

The cell means spread the N = 48 participants across the 6 cells, for n = 8. The main effect of practice difficulty spread those same N = 48 participants across just 3 conditions, for n = 16, Using n = 16 in the LSD Computator yields LSDmmd = 7.555. This would be used to compare the marginal means shown the "Descriptive Statistics" table (Easier = 50.6250, Harder = 66.6250, Same = 67.5000).

Please note: Because this design is non-orthogonal (has unequal n), this analysis is importantly different from the approach taken using the emmeans analysis above!

- The emmeans analysis tested and described the effect of practice difficulty after correcting practice difficulty for the effect of review attendance and the interaction. That is why it compared the estimated marginal means – estimated from the model.
- This approach compares the raw marginal means (without correction for the other effects in the model). The greater the non-orthogonality (unequal-n) of the design, the more these two analyses are likely to differ!

Which one to use? As you might expect, opinions differ, and the important things are to know what "your kind" expects and to be very clear which one you are presenting.

Describing the Main Effect of Review Attendance

/ emmeans tables (ar1y2n) compare (ar1y2n)

Estimates

Dependent Variable: testperf

ar1y2n	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Yes	68.333	2.134	64.026	72.641
No	54.444	2.320	49.762	59.127

Univariate Tests

Dependent Variable: testperf

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	2170.139	1	2170.139	19.406	.000
Error	4696.667	42	111.825		

The F tests the effect of ar1y2n. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

Pairwise Comparisons

Dependent Variable: testperf

(I) ar1y2n	(J) ar1y2n	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Yes	No	13.889*	3.153	.000	7.526	20.251
No	Yes	-13.889*	3.153	.000	-20.251	-7.526

Based on estimated marginal means

*. The mean difference is significant at the .050 level.

b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

As with the other main effect, you should notice that the means shown here are not the same as the marginal means from the “Descriptive Statistics” above (there 66.54 for Yes and 55.45 for No).

Also, the F-test for “ar1y2n” in the ANOVA table above and shown below (which match) are not comparing the data means shown in the “Descriptive Statistics” above.

Because there are unequal sample sizes among the design conditions, the main effects and the interaction are all collinear (nonorthogonal, or correlated). Thus, like all other multiple regressions, the model tests the unique contribution of each effect to the model, controlling for the other effects in the model.

So, in a factorial the main effects being tested are different than the raw data marginal means, the same as a multiple regression including quantitative variables will test a regression weight that is not the same as the bivariate correlation between a variable and the criterion!

The overall or main effect for Review Attendance is:

Review > No Review

However, we know from the pattern of the interaction that this is not descriptive for those in the Easier Practice condition.

This main effect must be communicated carefully, because it is potentially misleading.