2x2 2-Factor Between Groups ANOVA

The study examined the relationships of exam Review Attendance and Practice Difficulty with exam performance. Practice Difficulty was a 2-condition variable - practice problems were either easier than the exam problems (=1) or about the same difficulty as the exam problems (=2). Different sections of the course were randomly assigned to receive the two difficulty levels. The schedule showed the class meeting during which the exam review would occur & student's attendance was recorded (1= not attend, 2= attend). 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  AtndRev PractDif
```
← lists DV "by" IVs
order determines left-to-right ordering of IVs in the Descriptive Statistics table

```
  / method = sstype(3)
  / print descriptives
  / plot profile(PractDif * AtndRev)
  / design =  PractDif    AtndRev    PractDif*AtndRev.
```
← corrects each effect for all other effects
← get descriptive cell and marginal means
← get plot of cell means (x-axis * "separate lines")
← specify the design including the interaction that is automatically calculates from the IVs specified above)

Descriptive Statistics

Dependent Variable: TestPerf

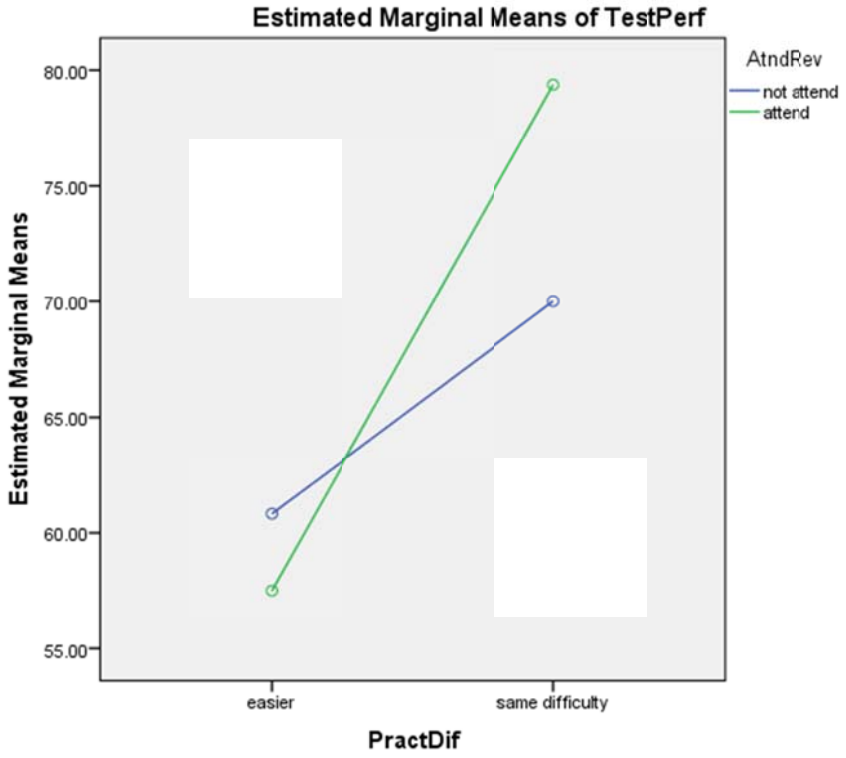
AtndRev	PractDif	Mean	Std. Deviation	N
not attend	easier	60.8333	7.92961	12
	same difficulty	70.0000	7.55929	8
	Total	64.5000	8.87041	20
attend	easier	57.5000	7.07107	8
	same difficulty	79.3750	7.71902	16
	Total	72.0833	12.84664	24
Total	easier	59.5000	7.59155	20
	same difficulty	76.2500	8.75388	24
	Total	68.6364	11.73167	44

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

((60.833 * 12) + (57.500 * 8)) / 20 = 59.500



From the means and the plots, it looks like performance is better after practice with same difficulty than with easier problems, and this effect is larger for those who attended the review.

Another way to describe the data pattern would be that for when using easier practice there is perhaps a small advantage to not attending the review, whereas when using the similar difficulty practices, there is a substantial advantage to attending the review.

Determine what effects are significant

Tests of Between-Subjects Effects

Dependent Variable: TestPerf

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	3582.765[a]	3	1194.255	20.455	.000
Intercept	181055.373	1	181055.373	3101.038	.000
PractDif	2434.320	1	2434.320	41.694	.000
AtndRev	92.215	1	92.215	1.579	.216
PractDif * AtndRev	408.004	1	408.004	6.988	.012
Error	2335.417	40	58.385		
Total	213200.000	44			
Corrected Total	5918.182	43			

a. R Squared = .605 (Adjusted R Squared = .576)

We have a significant effect for Practice Difficulty, no effect for Review Attendance and a significant interaction.

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 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 both those who practiced with similarly difficult 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 vs. 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 (AtndRev by PractDif) compare (PractDif)

← 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

AtndRev	PractDif	Mean	Std. Error
not attend	easier	60.833	2.206
	same difficulty	70.000	2.702
attend	easier	57.500	2.702
	same difficulty	79.375	1.910

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

Univariate Tests

Dependent Variable: TestPerf

AtndRev		Sum of Squares	df	Mean Square	F	Sig.
not attend	Contrast	403.333	1	403.333	6.908	.012
	Error	2335.417	40	58.385		
attend	Contrast	2552.083	1	2552.083	43.711	.000
	Error	2335.417	40	58.385		

Each F tests the simple effects of PractDif 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 2 Practice Difficulty conditions, the pairwise comparisons are redundant with the F-tests.

Not Attend $t^2 = (9.167 / 3.488)^2 = 6.908 = F$

Same $t^2 = (21.875 / 3.309)^2 = 43.711 = F$

Pairwise Comparisons

Dependent Variable: TestPerf

AtndRev	(I) PractDif	(J) PractDif	Mean Difference (I-J)	Std. Error	Sig.[b]
not attend	easier	same difficulty	-9.167*	3.488	.012
	same difficulty	easier	9.167*	3.488	.012
attend	easier	same difficulty	-21.875*	3.309	.000
	same difficulty	easier	21.875*	3.309	.000

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:

Not Attend
Easier **<** Same

Attend
Easier **<<** Same

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

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

/ emmeans tables (PractDif by AtndRev) compare (AtndRev)

← 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

PractDif	AtndRev	Mean	Std. Error
easier	not attend	60.833	2.206
	attend	57.500	2.702
same difficulty	not attend	70.000	2.702
	attend	79.375	1.910

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

PractDif		Sum of Squares	df	Mean Square	F	Sig.
easier	Contrast	53.333	1	53.333	.913	.345
	Error	2335.417	40	58.385		
same difficulty	Contrast	468.750	1	468.750	8.029	.007
	Error	2335.417	40	58.385		

Each F tests the simple effects of AtndRev 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.

Easier $t^2 = (3.333 / 3.488)^2 = .913 = F$

Same $t^2 = (9.375 / 3.309)^2 = 8.029 = F$

Pairwise Comparisons

Dependent Variable: TestPerf

PractDif	(I) AtndRev	(J) AtndRev	Mean Difference (I-J)	Std. Error	Sig.[b]
easier	not attend	attend	3.333	3.488	.345
	attend	not attend	-3.333	3.488	.345
same difficulty	not attend	attend	-9.375*	3.309	.007
	attend	not attend	9.375*	3.309	.007

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
Not Attend **=** Attend

Same Difficulty Practice
Not Attend **<** Attend

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

An Alternative Analysis of Cell Means to Describe Simple Effects and the Interaction

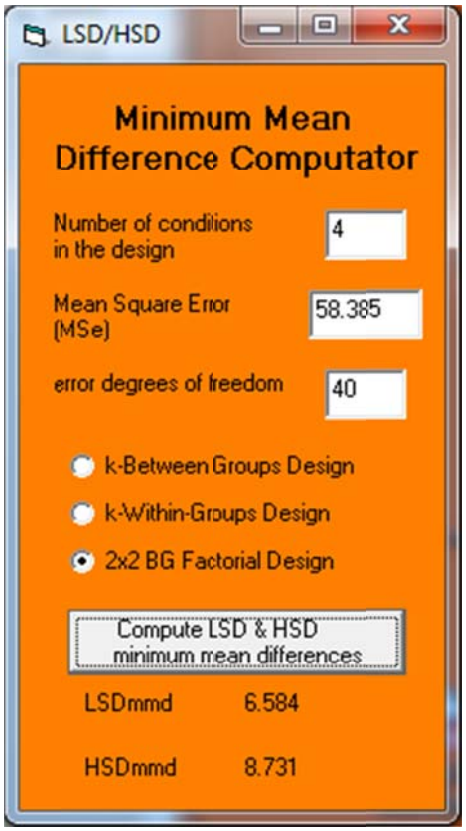
This is a BG model. Although all the F-tests and follow-up analyses are based on a single error term (MSe=58.385), the Standard Errors of the follow-ups vary with sample size.

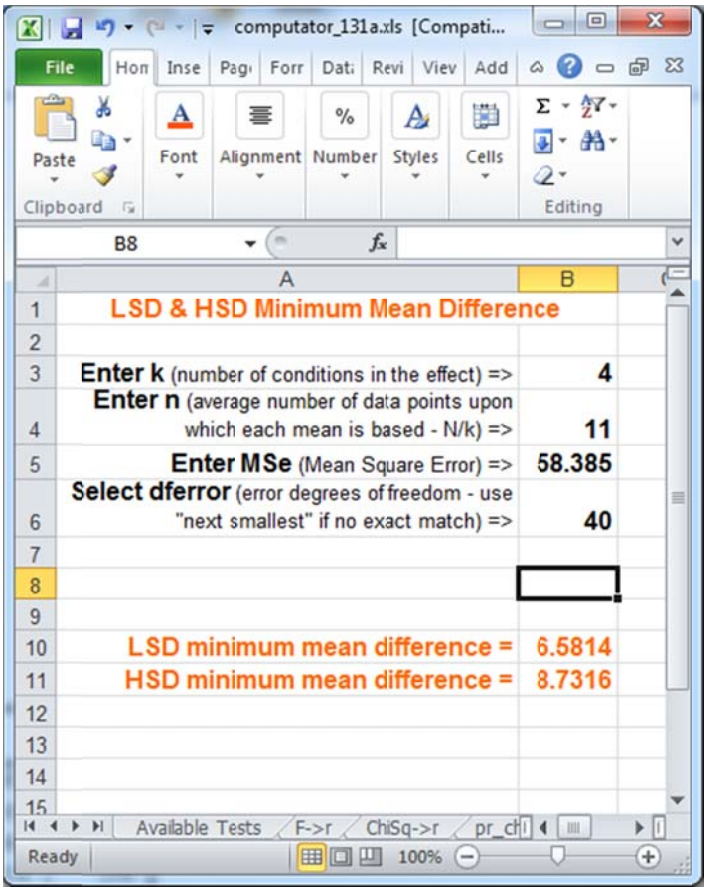
Why care? Because the follow-up analyses are based on a t-test (that isn't shown in the output, but how to compute it is shown above) that uses the standard error in the denominator. So, depending on whether the cells being compared have larger or smaller sample sizes, the standard error can be larger (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. Sometimes, with larger designs, 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.





Another Alternative Analysis of Cell Means to Describe Simple Effects and the Interaction

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.

```
sort cases by  PractDif .
```
← sorts the cases by the selection variable

```
temporary.
split file layered by  PractDif .
```
← specify that split command will only apply to the next analysis command

```
uninova testperf  by  AtndRev
  /design = AtndRev,
```
← splits the cases by the selection variable

← specify DV "by" IV (simple effect variable)

Tests of Between-Subjects Effects

Dependent Variable: TestPerf

PractDif	Source	Type III Sum of Squares	df	Mean Square	F	Sig.
easier	Corrected Model	53.333[a]	1	53.333	.922	.350
	Intercept	67213.333	1	67213.333	1161.446	.000
	AtndRev	53.333	1	53.333	.922	.350
	Error	1041.667	18	57.870		
	Total	71900.000	20			
	Corrected Total	1095.000	19			
same difficulty	Corrected Model	468.750[b]	1	468.750	7.971	.010
	Intercept	119002.083	1	119002.083	2023.610	.000
	AtndRev	468.750	1	468.750	7.971	.010
	Error	1293.750	22	58.807		
	Total	141300.000	24			
	Corrected Total	1762.500	23			

a. R Squared = .049 (Adjusted R Squared = -.004)

b. R Squared = .266 (Adjusted R Squared = .233)

The SS effect (AtndRev) are the same as from the EMMEANS analyses above. Each compares the same cell means

The SS Error are different from the EMMEANS analyses above. These are based on data from two cells, while EMMEANS were based on data from all four cells.

The df-error are different from the EMMEANs analyses above. These are based on n from the two cells being compared, while EMMEANS were based on n from all four cells.

The MSerror are different, because both the SSerror and df-error are different.

The F-values and p-values are different.

Tests of Between-Subjects Effects

Dependent Variable: TestPerf

AtndRev	Source	Type III Sum of Squares	df	Mean Square	F	Sig.
not attend	Corrected Model	403.333[a]	1	403.333	6.650	.019
	Intercept	82163.333	1	82163.333	1354.754	.000
	PractDif	403.333	1	403.333	6.650	.019
	Error	1091.667	18	60.648		
	Total	84700.000	20			
	Corrected Total	1495.000	19			
attend	Corrected Model	2552.083[b]	1	2552.083	45.142	.000
	Intercept	99918.750	1	99918.750	1767.407	.000
	PractDif	2552.083	1	2552.083	45.142	.000
	Error	1243.750	22	56.534		
	Total	128500.000	24			
	Corrected Total	3795.833	23			

a. R Squared = .270 (Adjusted R Squared = .229)

b. R Squared = .672 (Adjusted R Squared = .657)

Here is the syntax to get the simple effects of practice difficulty for each review attendance condition.

```
sort cases by   AtndRev.

temporary.
split file layered by  AtndRev.

uninova testperf  by  PractDif
  /design = PractDif.
```

Describing the Main Effect of Review Attendance

/ emmeans tables (AtndRev) compare (AtndRev)

Estimates

Dependent Variable: TestPerf

AtndRev	Mean	Std. Error
not attend	65.417	1.744
attend	68.438	1.654

Univariate Tests

Dependent Variable: TestPerf

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	92.215	1	92.215	1.579	.216
Error	2335.417	40	58.385		

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

Pairwise Comparisons

Dependent Variable: TestPerf

(I) AtndRev	(J) AtndRev	Mean Difference (I-J)	Std. Error	Sig.[a]
not attend	attend	-3.021	2.404	.216
attend	not attend	3.021	2.404	.216

Based on estimated marginal means

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

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

Easier Practice Not Attend **=** Attend

Same Difficulty Practice Not Attend **<** Attend

You should notice that the means shown here are not the same as the marginal means from the "Descriptive Statistics" above (there 64.5 for Not Attend and 72.08 for Attend).

Also, the F-test for "AtndRev" 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 multivariate analyses using Type III SS, the model tests the unique contribution of each effect to the model, controlling for the other effects in the model.

So, in a factorial using Type III SS, 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:

Attend **=** Not Attend

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

Describing the Main Effect of Practice Difficulty

/ emmeans tables (PractDif) compare (PractDif)

Estimates

Dependent Variable: TestPerf

PractDif	Mean	Std. Error
easier	59.167	1.744
same difficulty	74.688	1.654

Univariate Tests

Dependent Variable: testperf

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	2210.278	2	1105.139	9.883	.000
Error	4695.667	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) PractDif	(J) PractDif	Mean Difference (I-J)	Std. Error	Sig.[b]
easier	same difficulty	-15.521*	2.404	.000
same difficulty	easier	15.521*	2.404	.000

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).

Again, you should notice the means shown here are not the same as the marginal means from the "Descriptive Statistics" above (59.5 for Easier and 76.25 for Same).

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

Also, we know from the pattern of the interaction that this is descriptive.

Not Attend Easier **<** Same

Attend Easier **<<** Same