2xk 2-Factor Mixed Groups ANOVA

The purpose of this study was to examine the Practice Difficulty on exam performance for each of the four exam given during the semester. 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 dependent variable was performance on each of the four examinations given during the semester.

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

```
glm TestPerf1 TestPerf2 TestPerf3 TestPerf4      ← lists DV -- list each variable that is DV for one of the IV conditions
     BY PractDif                                 ← "by"  IV
  /wsfactor=AllTests 4                           ← give a name to the WG IV (can't match any variable name)
  /method=sstype(3)                              ← corrects each effect for all other effects
  /print=descriptive                             ← get descriptive cell and marginal means
  /plot=profile(AllTests*PractDif)               ← get plot of cell means (x-axis * "separate lines" )
  /wsdesign=AllTests                             ← identifies WG IV
  /design=PractDif.                              ← identifies BG IV (interactions are automatically generarated)
```

Descriptive Statistics

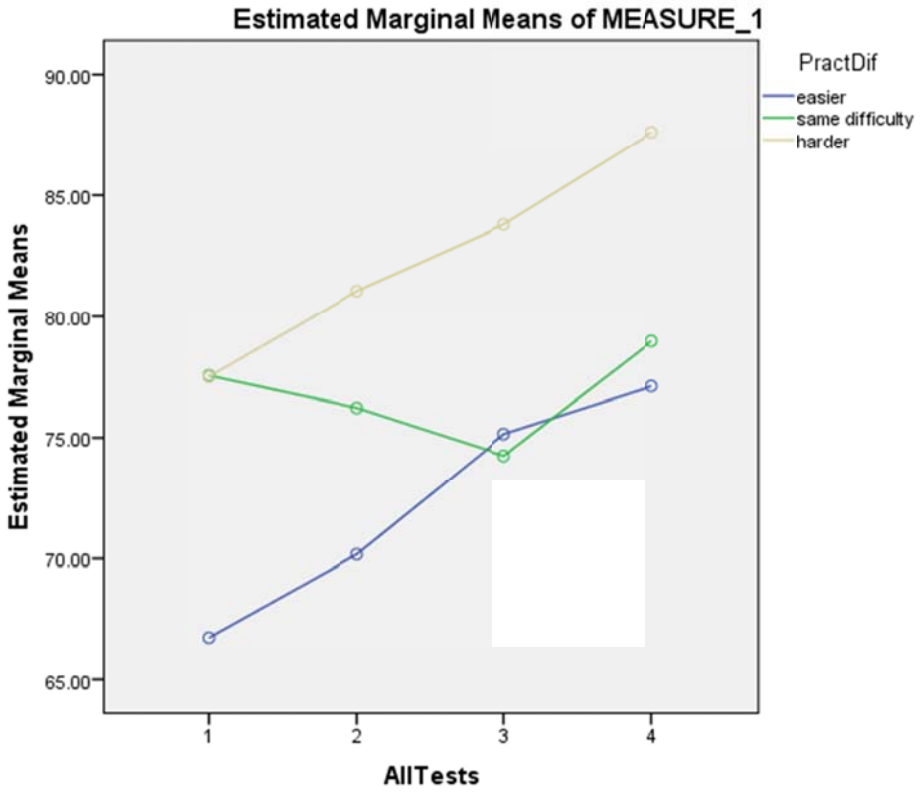
	PractDif	Mean	Std. Deviation	N
TestPerf1	easier	66.7461	9.11793	20
	same difficulty	77.5523	9.44029	24
	harder	77.5091	10.84393	25
	Total	74.4044	10.91633	69
TestPerf2	easier	70.1782	9.33466	20
	same difficulty	76.2272	6.82222	24
	harder	81.0549	7.54833	25
	Total	76.2230	8.91792	69
TestPerf3	easier	75.1510	7.52676	20
	same difficulty	74.2464	9.43451	24
	harder	83.7984	9.08987	25
	Total	77.9694	9.73713	69
TestPerf4	easier	77.1216	8.76037	20
	same difficulty	78.9812	7.63787	24
	harder	87.6087	6.61513	25
	Total	81.5681	8.84380	69

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 TestPerf1 is
((66.7461*20) + (77.5523*24) + (77.5091*25)) / 69 = 74.4044

Notice that the marginal means for the BG main effect are not given (more below!).



From the means and the plots, it looks like relative performance of the exam difficulty groups changed across the four exams. Harder and same difficulty practice both outperformed easier on Exam 1. For Exam 2 harder outperformed same, which outperformed easier. While for both Exams 3 and 4, harder outperformed same and easier.

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
AllTests	Sphericity Assumed	1991.778	3	663.926	9.152	.000
	Greenhouse-Geisser	1991.778	2.675	744.499	9.152	.000
	Huynh-Feldt	1991.778	2.884	690.747	9.152	.000
	Lower-bound	1991.778	1.000	1991.778	9.152	.004
AllTests * PractDif	Sphericity Assumed	1065.582	6	177.597	2.448	.026
	Greenhouse-Geisser	1065.582	5.351	199.150	2.448	.032
	Huynh-Feldt	1065.582	5.767	184.771	2.448	.028
	Lower-bound	1065.582	2.000	532.791	2.448	.094
Error(AllTests)	Sphericity Assumed	14364.251	198	72.547		
	Greenhouse-Geisser	14364.251	176.572	81.351		
	Huynh-Feldt	14364.251	190.312	75.477		
	Lower-bound	14364.251	66.000	217.640		

The ANOVA results are given in two summary tables.

The WG main effect and the interaction are shown in one table, with multiple F-tests.

The “Sphericity Assumed” is the traditional approach. The others are various attempts to correct the p-value for departures from the assumptions of the model.

Both the inteaction and the main effect of test-retest are significant.

Tests of Between-Subjects Effects

Measure: MEASURE_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	1628734.247	1	1628734.247	20923.318	.000
PractDif	4709.893	2	2354.947	30.253	.000
Error	5137.639	66	77.843		

The between groups main effect of practice difficulty is also significant.

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.
- The an average of the four tests seems a reasonable aggregate to represent the semester's performance, so this main effect might be interesting, especially if the main effect pattern is descriptive for a majority of the exams.

Four Exams

- The marginal means are of dubious value, because it is unlikely that a group of students will practice for an exam with a variety of differently difficult practice problems. 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, "The purpose of this study was to examine the Practice Difficulty on exam performance for each of the four exam given during the semester. ." 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 for each of the four exams.

Obtaining and describing the pairwise simple effects of Practice Difficulty for each Exam

/emmeans=tables(AllTests*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

Measure: MEASURE_1

AllTests	PractDif	Mean	Std. Error
1	easier	66.746	2.211
	same difficulty	77.552	2.018
	harder	77.509	1.977
2	easier	70.178	1.761
	same difficulty	76.227	1.608
	harder	81.055	1.575
3	easier	75.151	1.967
	same difficulty	74.246	1.796
	harder	83.798	1.759
4	easier	77.122	1.708
	same difficulty	78.981	1.559
	harder	87.609	1.528

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

Remember that this is the set of BG simple effects in this MG factorial. So, the simple F-tests and pairwise comparisons are computed using a BG error term. Notice that the dferror (66) for the simple effect F-tests match those from the test of Practice Difficulty BG main effect test in the omnibus ANOVA above.

.

Univariate Tests

Measure: MEASURE_1

AllTests		Sum of Squares	df	Mean Square	F	Sig.
1	Contrast	1651.797	2	825.899	8.449	.001
	Error	6451.513	66	97.750		
2	Contrast	1314.479	2	657.239	10.597	.000
	Error	4093.518	66	62.023		
3	Contrast	1340.951	2	670.475	8.665	.000
	Error	5106.637	66	77.373		
4	Contrast	1468.248	2	734.124	12.584	.000
	Error	3850.222	66	58.337		

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 of the four exams.

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

The pairwise comparisons for these simple effects are shown on the next page.

The pairwise effects describing the pattern of the interaction are:

	Easier v Same	Easier v Harder	Same v Harder
Test1	**<**	**<**	**=**
Test2	**<**	**<**	**<**
Test3	**=**	**<**	**<**
Test4	**=**	**<**	**<**

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

Pairwise Comparisons

Measure: MEASURE_1

AllTests	(I) PractDif	(J) PractDif	Mean Difference (I-J)	Std. Error	Sig.[b]
1	easier	same difficulty	-10.806*	2.993	.001
		harder	-10.763*	2.966	.001
	same difficulty	easier	10.806*	2.993	.001
		harder	.043	2.825	.988
	harder	easier	10.763*	2.966	.001
		same difficulty	-.043	2.825	.988
2	easier	same difficulty	-6.049*	2.384	.014
		harder	-10.877*	2.363	.000
	same difficulty	easier	6.049*	2.384	.014
		harder	-4.828*	2.251	.036
	harder	easier	10.877*	2.363	.000
		same difficulty	4.828*	2.251	.036
3	easier	same difficulty	.905	2.663	.735
		harder	-8.647*	2.639	.002
	same difficulty	easier	-.905	2.663	.735
		harder	-9.552*	2.514	.000
	harder	easier	8.647*	2.639	.002
		same difficulty	9.552*	2.514	.000
4	easier	same difficulty	-1.860	2.312	.424
		harder	-10.487*	2.291	.000
	same difficulty	easier	1.860	2.312	.424
		harder	-8.627*	2.183	.000
	harder	easier	10.487*	2.291	.000
		same difficulty	8.627*	2.183	.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).

If you are asked the t-value for each pairwise comparison is…

t = Mean Difference / Std. Error

and the df = the dferror for testing the BG main effect in the overall ANOVA (66).

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

/emmeans=tables(PractDif*AllTests) compare(AllTests)

← 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

Measure: MEASURE_1

PractDif	AlTests	Mean	Std. Error
easier	1	66.746	2.211
	2	70.178	1.761
	3	75.151	1.967
	4	77.122	1.708
same difficulty	1	77.552	2.018
	2	76.227	1.608
	3	74.246	1.796
	4	78.981	1.559
harder	1	77.509	1.977
	2	81.055	1.575
	3	83.798	1.759
	4	87.609	1.528

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

.

The F-tests SPSS provides for these within-subjects simple effects are based on a somewhat different "multivariate" approach to comparing the effect means. Since the pairwise comparisons provide the important portion of the analysis, we will focus on those.

If you are asked for the t-value corresponding to a p-value for any pairwise comparison…

t = Mean Difference / Std. Error
df = dferror from the WG main effect and Interaction F-tests

The pairwise comparisons for these simple effects are shown on the next page.

The pattern of the interaction is:

	Test1 v Test2	Test1 v Test3	Test2 v Test4	Test2 v Test3	Test2 v Test4	Test3 v Test4
Easier	=	<	<	=	<	=
Same Difficulty	=	=	=	=	=	<
Harder	=	=	<	=	<	=

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

Pairwise Comparisons

Measure: MEASURE_1

PractDif	(I) AllTests	(J) AllTests	Mean Difference (I-J)	Std. Error	Sig.[b]
easier	1	2	-3.432	2.661	.202
		3	-8.405*	3.287	.013
		4	-10.375*	2.599	.000
	2	1	3.432	2.661	.202
		3	-4.973	2.650	.065
		4	-6.943*	2.311	.004
	3	1	8.405*	3.287	.013
		2	4.973	2.650	.065
		4	-1.971	2.554	.443
	4	1	10.375*	2.599	.000
		2	6.943*	2.311	.004
		3	1.971	2.554	.443
same difficulty	1	2	1.325	2.429	.587
		3	3.306	3.001	.275
		4	-1.429	2.372	.549
	2	1	-1.325	2.429	.587
		3	1.981	2.419	.416
		4	-2.754	2.110	.196
	3	1	-3.306	3.001	.275
		2	-1.981	2.419	.416
		4	-4.735*	2.332	.046
	4	1	1.429	2.372	.549
		2	2.754	2.110	.196
		3	4.735*	2.332	.046
harder	1	2	-3.546	2.380	.141
		3	-6.289*	2.940	.036
		4	-10.100*	2.324	.000
	2	1	3.546	2.380	.141
		3	-2.743	2.370	.251
		4	-6.554*	2.067	.002
	3	1	6.289*	2.940	.036
		2	2.743	2.370	.251
		4	-3.810	2.285	.100
	4	1	10.100*	2.324	.000
		2	6.554*	2.067	.002
		3	3.810	2.285	.100

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).

Please note that the Std Errors used for the WG pairwise comparisons up above are substantially smaller than the Std Error used for these BG pairwise comparisons. See the discussion in the next section!

Alternative Analysis of Cell Means

This is a MG model. The WG main effect and interaction F-tests are based on one error term and the BG main effect is based on another error term. However, the follow-up analyses are each based on a specific error term, and 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.

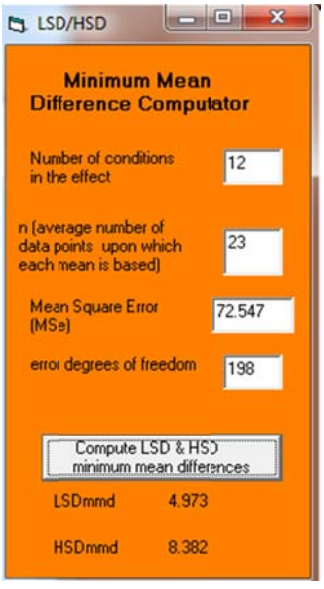
Another issue with mixed groups designs involves the choice of the error term to use to test the pairwise simple effects. In a mixed factorial, the interaction is tested as a within-groups effect, using the within-groups error term, generally leading to a more powerful test than would a corresponding comparison using a between groups model and error term.

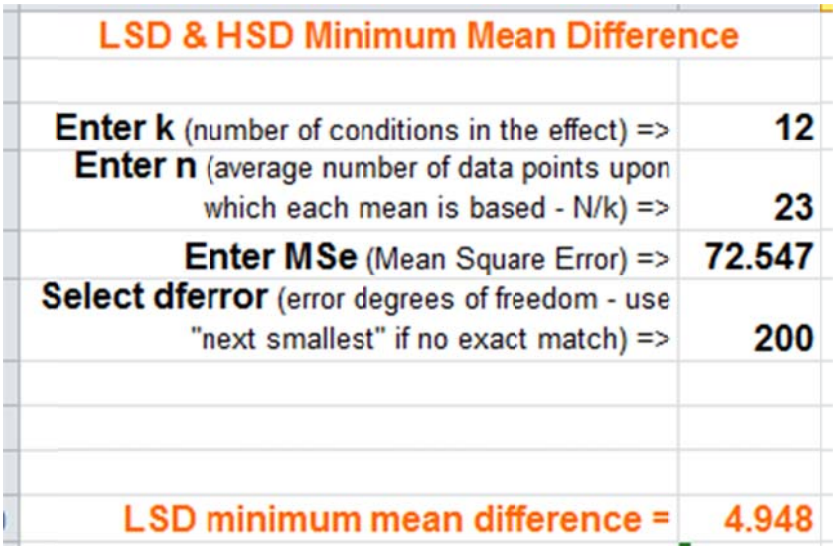
SPSS uses a BG error term to compare the BG simple effect within the MG interaction e.g., Easier vs. Same Difficulty, Std Errors = 2.497 & 2.535) and it uses a WG error term to compare the WG simple effect with the interaction (e.g., Test vs. Retest, Std Errors = .734 & .670). The WG error terms are smaller and the WG pairwise comparisons are consequently more powerful, than for the BG simple effect. One possible consequence that the examination of the WG comparisons provides evidence of an interaction pattern, while the BG comparisons do not, simply because of differential power! This has led some to recommend always examining significant MG interactions using the WG pairwise comparisons. While solid statistical advice, what are we to do when the BG IV is the "primary" variable in the factorial, and our intent was to describe how this effect is moderated by the WG IV?

An alternative is to use this WG error term that was used to test the interaction 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 WG error term approach is that since the proper error term to test the interaction, it is also the proper error term to compare the associated cell means to explicate the pattern of the interaction.

Why do people who like this approach like it?

1. It is based on the same estimate of variability, but larger sample size and the WG error term, and, so, uses a smaller standard error than the pairwise error term approach use by SPSS, especially when comparing the BG simple effects. 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.





Describing the BG Main Effect of Practice Difficulty

/emmeans=tables(PractDif) compare(PractDif)

Estimates

Measure: MEASURE_1

PractDif	Mean	Std. Error
easier	72.299	.986
same difficulty	76.752	.900
harder	82.493	.882

Univariate Tests

Measure: MEASURE_1

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	1177.473	2	588.737	30.253	.000
Error	1284.410	66	19.461		

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

You should notice that the marginal means for this main effect were not given in the "Descriptives" table at the beginning of the analysis output!

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

Pairwise Comparisons

Measure: MEASURE_1

(I) PractDif	(J) PractDif	Mean Difference (I-J)	Std. Error	Sig.[b]
easier	same difficulty	-4.453*	1.336	.001
	harder	-10.194*	1.323	.000
same difficulty	easier	4.453*	1.336	.001
	harder	-5.741*	1.261	.000
harder	easier	10.194*	1.323	.000
	same difficulty	5.741*	1.261	.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).

However, we know from the pattern of the interaction that this is only descriptive for Test2 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 "n" for the LSD computation is the number of data points, not the number of cases, each marginal mean is based on. For this design, with 4 WG conditions, and 69 cases spread across 3 BG conditions, this would be 4 * 23 = 93.

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 WG Main Effect of Tests

/emmeans=tables(AllTests) compare(AllTests)

Estimates

Measure: MEASURE_1

AllTests	Mean	Std. Error
1	73.936	1.196
2	75.820	.953
3	77.732	1.064
4	81.237	.924

You should notice that the means shown here are not the same as the marginal means from the "Descriptive Statistics" above (Test1 = 74.40, Test2 = 76.22, Test3 = 77.97 & Test4 = 81.57)

The F-tests SPSS provides for these within-subjects simple effects are based on a somewhat different "multivariate" approach to comparing the effect means. Since the pairwise comparisons provide the important portion of the analysis, we will focus on those.

If you are asked for the t-value corresponding to a p-value for any pairwise comparison…

t = Mean Difference / Std. Error
df = dferror from the interaction F-test

Pairwise Comparisons

Measure: MEASURE_1

(I) AllTests	(J) AllTests	Mean Difference (I-J)	Std. Error	Sig.[b]
1	2	-1.884	1.439	.195
	3	-3.796*	1.778	.036
	4	-7.301*	1.406	.000
2	1	1.884	1.439	.195
	3	-1.912	1.433	.187
	4	-5.417*	1.250	.000
3	1	3.796*	1.778	.036
	2	1.912	1.433	.187
	4	-3.505*	1.382	.014
4	1	7.301*	1.406	.000
	2	5.417*	1.250	.000
	3	3.505*	1.382	.014

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).

Also, the F-test for "AllTests" in the ANOVA table above and the pairwise comparison 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 pattern of the Main Effect is:

Test1 v Test2	Test1 v Test3	Test2 v Test4	Test2 v Test3	Test2 v Test4	Test3 v Test4
=	<	<	=	<	<

However, we know from the pattern of the interaction that this main effect must be communicated carefully, because it is potentially misleading. There is no practice difficulty condition for which there is this pattern of test effects.