2x2 2-Factor Mixed Groups ANOVA

The study examined the relationships of exam Practice Difficulty with exam performance in a Test-Retest format. We wanted to know if performance changes from the test to the retest were different for those in the two practice difficulty conditions. 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. Each person took the examination, received their grade and feedback and retook the examination (the examinations were constructed and piloted to ensure comparable difficulty). Each student completed 5 practice assignments of the assigned type, to a minimum performance criterion, before each exam. The dependent variable was performance on each 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

```
glm TestPerf reTestPerf                  ← lists DV -- list each variable that is the DV for one of the IV conditions
BY PractDif                              ← "by"  IV
  /wsfactor=Test_reTest 2                ← 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(Test_reTest*PractDif)    ← get plot of cell means (x-axis * "separate lines" )
  /wsdesign=Test_reTest                  ← identifies WG IV
  /design=PractDif.                      ← identifies BG IV (interactions are automatically generarated)
```

Descriptive Statistics

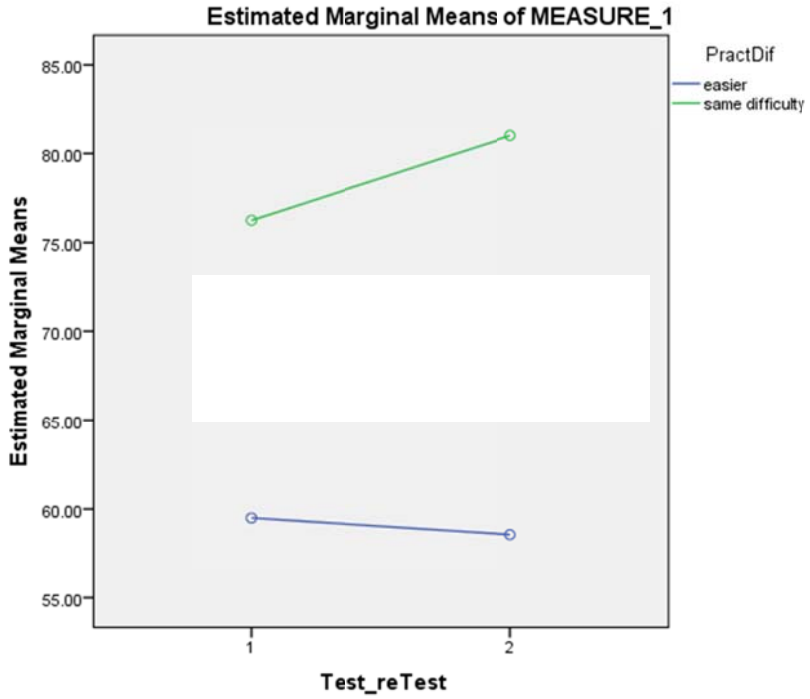
	PractDif	Mean	Std. Deviation	N
TestPerf	easier	59.5000	7.59155	20
	same difficulty	76.2500	8.75388	24
	Total	68.6364	11.73167	44
reTestPerf	easier	58.5613	7.18817	20
	same difficulty	81.0100	9.23894	24
	Total	70.8060	14.01206	44

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
((59.500 * 20) + (76.250 * 24)) / 44 = 68.6364

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



From the means and the plots, it looks like retest performance was equivalent when using easier practice, but improved for those in the same difficulty condition.

Another way to describe data pattern would be that those in the same difficulty condition performed better on both the test and the retest, with larger difference on the retest.

Determine what effects are significant

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Test_reTest	Sphericity Assumed	79.647	1	79.647	14.784	.000
	Greenhouse-Geisser	79.647	1.000	79.647	14.784	.000
	Huynh-Feldt	79.647	1.000	79.647	14.784	.000
	Lower-bound	79.647	1.000	79.647	14.784	.000
Test_reTest * PractDif	Sphericity Assumed	177.136	1	177.136	32.879	.000
	Greenhouse-Geisser	177.136	1.000	177.136	32.879	.000
	Huynh-Feldt	177.136	1.000	177.136	32.879	.000
	Lower-bound	177.136	1.000	177.136	32.879	.000
Error(Test_reTest)	Sphericity Assumed	226.277	42	5.388		
	Greenhouse-Geisser	226.277	42.000	5.388		
	Huynh-Feldt	226.277	42.000	5.388		
	Lower-bound	226.277	42.000	5.388		

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. With 2-groups, the forms will agree.

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	413464.340	1	413464.340	3114.227	.000
PractDif	8381.109	1	8381.109	63.127	.000
Error	5576.183	42	132.766		

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.
- More importantly, it is unclear what population is represented by an average the test and the retest
- So, this main effect is only likely to be interesting if that main effect is descriptive, and so, it describes the comparison of the groups for both the test and the retest.

Test_reTest

- This is a classic WG or repeated measure IV
- 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 performance changes from the test to the retest were different for those in the two practice difficulty conditions 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 test vs. retest and then examine how this simple effect is different those who practiced with the similar difficulty versus easier problems.

Obtaining and describing the pairwise simple effects of Test-Retest for each level of Practice Difficulty

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

← 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

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

Estimates

Measure: MEASURE_1

PractDif	Test_reTest	Mean	Std. Error
easier	1	59.500	1.844
	2	58.561	1.872
same difficulty	1	76.250	1.684
	2	81.010	1.709

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

Pairwise Comparisons

Measure: MEASURE_1

PractDif	(I) Test_reTest	(J) Test_reTest	Mean Difference (I-J)	Std. Error	Sig.[b]
easier	1	2	.939	.734	.208
	2	1	-.939	.734	.208
same difficulty	1	2	-4.760*	.670	.000
	2	1	4.760*	.670	.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:

Easier
Test = Retest

Same Difficulty
Test < Retest

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

Obtaining and describing the pairwise simple effects of Practice Difficulty for Tests and Retest

/emmeans=tables(Test_reTest*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

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

Estimates

Measure: MEASURE_1

Test_reTest	PractDif	Mean	Std. Error
1	easier	59.500	1.844
	same difficulty	76.250	1.684
2	easier	58.561	1.872
	same difficulty	81.010	1.709

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

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 (42) 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

Test_reTest		Sum of Squares	df	Mean Square	F	Sig.
1	Contrast	3060.682	1	3060.682	44.986	.000
	Error	2857.500	42	68.036		
2	Contrast	5497.563	1	5497.563	78.404	.000
	Error	2944.960	42	70.118		

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.

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

Easier $t^2 = (16.75 / 2.497)^2 = 44.986 = F$

Same $t^2 = (22.449 / 2.535)^2 = 78.404 = F$

Pairwise Comparisons

Measure: MEASURE_1

Test_reTest	(I) PractDif	(J) PractDif	Mean Difference (I-J)	Std. Error	Sig.[b]
1	easier	same difficulty	-16.750*	2.497	.000
	same difficulty	easier	16.750*	2.497	.000
2	easier	same difficulty	-22.449*	2.535	.000
	same difficulty	easier	22.449*	2.535	.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:

Test
Easier < Same Difficulty

Retest
Easier < Same Difficulty

We know these two simple effects are of different size, because the interaction is significant!

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

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!

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

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.

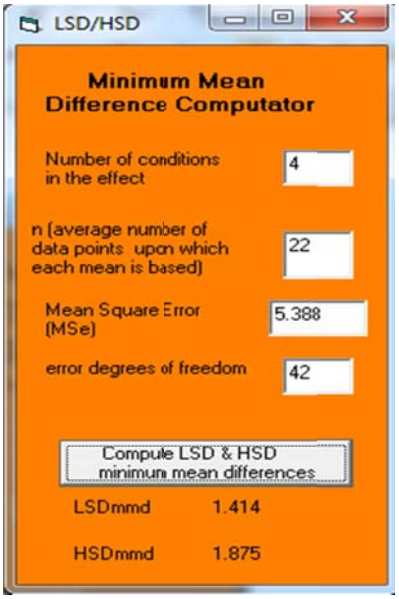
Another issue with mixed groups designs involves the choice of the error term to use to test the pairwise simple effects. In a mixed factorial, the interaction is tested as a within-groups effect, using the within-groups error term, generally leading to a more powerful test than would a corresponding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

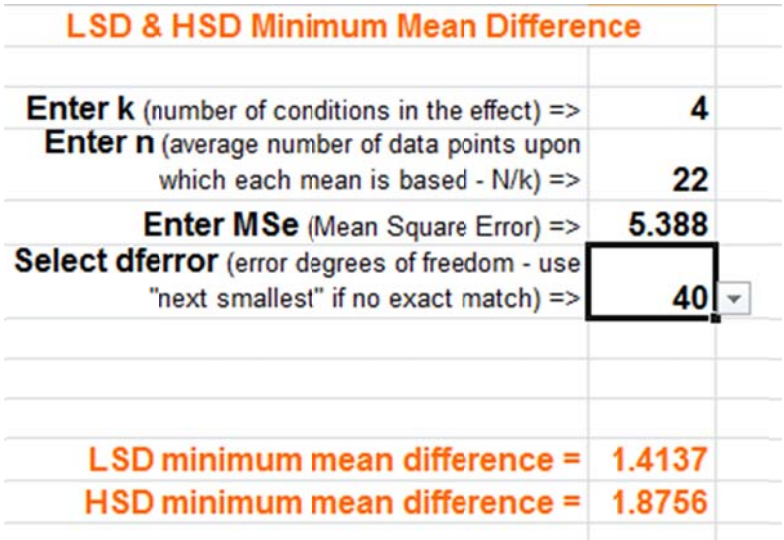
SPSS uses a BG error term to compare the BG simple effect within the MG interaction e.g., Easier vs. Same Difficulty, Std Errors = 2.497 & 2.535) and it uses a WG error term to compare the WG simple effect with the interaction (e.g., Test vs. Retest, Std Errors = .734 & .670). The WG error terms are 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 WG Main Effect of Test-Retest

/emmeans=tables(Test_reTest) compare(Test_reTest)

Estimates

Measure: MEASURE_1

Test_reTest	Mean	Std. Error
1	67.875	1.249
2	69.786	1.268

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) Test_reTest	(J) Test_reTest	Mean Difference (I-J)	Std. Error	Sig.[b]
1	2	-1.911*	.497	.000
2	1	1.911*	.497	.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).

You should notice that the means shown here are not the same as the marginal means from the "Descriptive Statistics" above (there 68.6364 for Test and 70.8060 for Retest).

Also, the F-test for "Test_reTest" 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 overall or main effect for Test – Retest is:

Test < Retest

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

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

Easier Test = Retest

Same Difficulty Test < Retest

Describing the BG Main Effect of Practice Difficulty

/emmeans=tables(PractDif) compare(PractDif)

Estimates

Measure: MEASURE_1

PractDif	Mean	Std. Error
easier	59.031	1.822
same difficulty	78.630	1.663

Univariate Tests

Measure: MEASURE_1

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	4190.554	1	4190.554	63.127	.000
Error	2788.092	42	66.383		

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

Pairwise Comparisons

Measure: MEASURE_1

(I) PractDif	(J) PractDif	Mean Difference (I-J)	Std. Error	Sig.[b]
easier	same difficulty	-19.599*	2.467	.000
same difficulty	easier	19.599*	2.467	.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).

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 < Same Difficulty

We know from the interaction pattern that the main effect of Practice Difficulty will be **descriptive!**

Test Easier < Same Difficulty

Retest Easier < Same Difficulty