Dependent Sample t-tests & ANOVA (within-subjects) & effect size

By Nicole Rose & Xueli Tang

About the topic

- **ANOVAS:** belong to the F family of tests and they are parametric tests normally used for determining if, based on their means, >2 samples are statistically significantly different (Miguel, 2021).
- **Dependent sample t-tests:** it is used to compare the differences in means of two related groups (* the same participants will be involved in both groups).
- **Different terms used for dependent sample t-tests:** within-subjects measures, repeated measures, paired samples, before and after measures, matched pairs...
- **Note the difference:** whether 3 or more related groups are measured?

When to use a dependent sample t-test with example?

- EG. To investigate whether there is a difference between students' performance in math class before and after a 2-month intensive training program.

- differences before & after (values)

- 1 IV (2 levels with and without the training) & 1DV

- Same participants involved in both conditions;

What is within-subjects ANOVA?

- **Different from between-subjects ANOVA!**
- **Within-subjects ANOVA= Repeated measures ANOVA :**

One-way ANOVA (for 3 or more related groups);

Extension of dependent sample t-test;

To discover overall differences;

When to use within-subjects ANOVA with example?

- Eg. Investigation on the effect of a 2-month vocabulary training program on students' reading comprehension at 3 different time points.
- 1 IV (2levels with & without the training) & 1DV;
- Same participants involved in both conditions, but differences in 3 time points;
- **To find out:** changes in mean scores over various time points or differences in mean scores over various conditions

What are the assumptions to be considered?

Dependent sample t-test & within-subjects ANOVA:

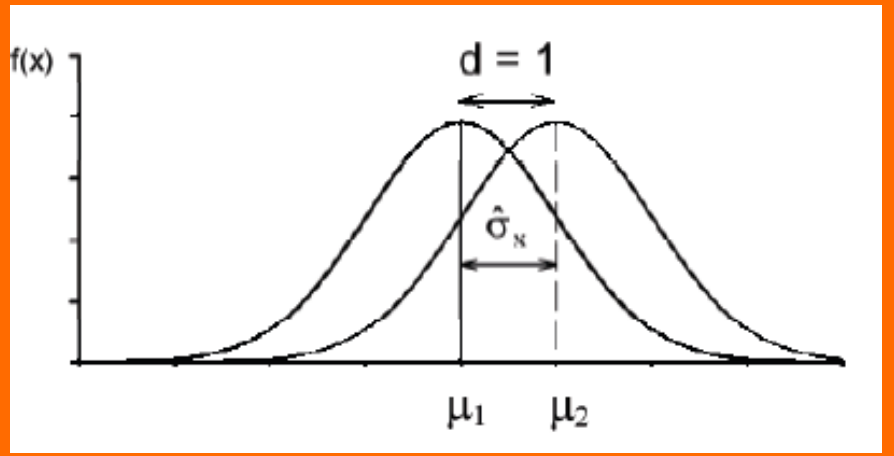
- *DV continuous (interval or ratio) & IV categorical (nominal or ordinal);*
- *Normally distributed differences in DV & no outliers;*
- *Correlated groups are required;*
- *Sphericity to be assumed. (The variances of differences between related groups are equal)*

(Same parametric assumptions as mentioned for topic 9 "ANOVA between-subjects design".)

Question to think?

- **Investigation on the effects of video games on learners' math scores at 3 time points (pre, one-month & post).**
- IV & possible factor levels?
- DV?
- Dependent sample t-test or a within-subjects ANOVA? (If possible, give a reason)

Effect Size



What is the effect size?

- Quantification of the size of the difference between two group means
- Quantification of the size of association between variables

Why do we need the effect size?

- Whereas a p-value indicates if an intervention works, the effect size indicates how much an intervention works, independent of sample size
- Effect sizes are standardized

Questions to think about:

- Why is it an issue that the p-value is dependent on the sample size? (As one of the reasons why we need effect size)
- What is the advantage of effect sizes being standardized?

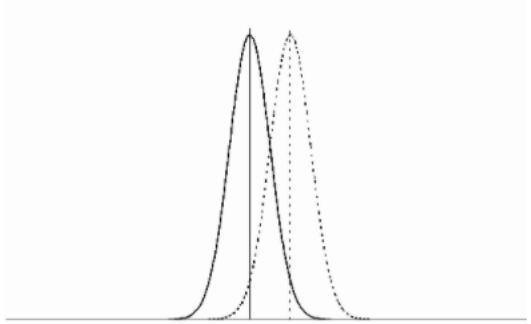
What is the effect size?

Questions to think about:

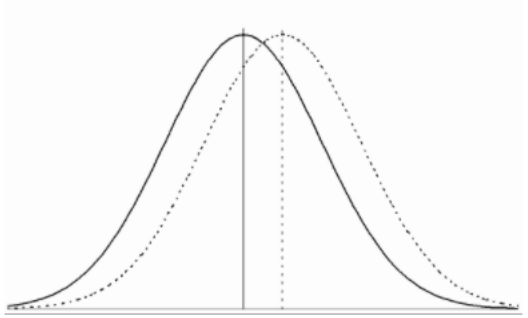
- Why is it an issue that the p-value is dependent on the sample size?
  - When the sample size is small, strong and important effects can be non-significant (Type II Error is made)
  - When the sample size is large, even trivial effects can have significant p-values

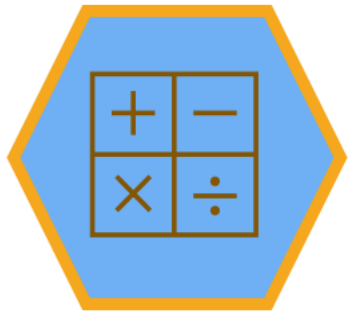
- What is the advantage of effect sizes being standardized?
  - We can quantitatively compare the results of studies conducted in different settings

Cohen's *d*



vs.





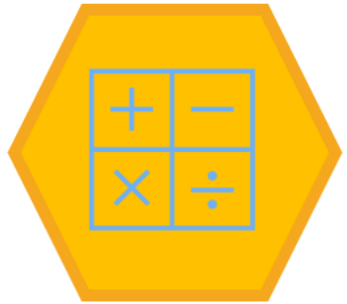
➡ the graphs differ regarding their variance

- $Effect\ Size = value\ of\ standardized\ distance\ between\ two\ means = d = \frac{\mu_{experimental} - \mu_{control}}{\sigma}$
- *d* = 1 indicates the means differ by one standard deviation

  ➡ e.g *d* = 0.8 means that on average, an object of the experimental group scores 0.8 standard deviations higher than the average person of the control group

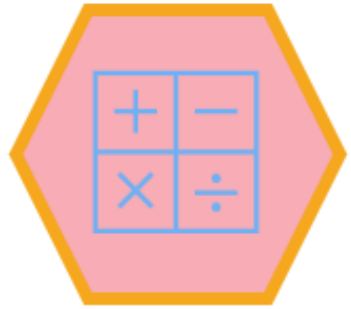
- Cohen's *d* should accompany the results of **t-tests**, especially if results are significant

η^2 (eta squared)



- Eta squared is the proportion of variance accounted for by main effects or interaction effects in **ANOVA**
- The sum of squares is a measure of how much an entire set of data varies around a mean
- $\eta^2 = SS_{between\text{-}groups} / SS_{total}$ *between-subjects ANOVA*
- $\eta^2 = SS_{conditions} / SS_{total}$ *repeated measures ANOVA*
  - $SS_{between\text{-}groups}$ or $SS_{conditions}$ is the sum of squares of the effect you are looking at
  - SS_{total} is the sum of squares of all effects, errors and interactions: it tells us how much variation there is in the dependent variable
- η^2 is additive and can never exceed 1; i.e. one cannot account for more than 100% of the variance

η^2 (eta squared) – Example



- For example, we are studying people's happiness self-rated on a 100 point scale. The considered factors are participants' gender (male vs. female) and their employment status (employed vs. unemployed vs. part-time employed.
- After performing ANOVA, we get the following results:
  - Total SS = 62.29
  - Gender SS = 13.24
  - Employment Status SS = 19.58
- Dividing each SS by the Total SS gives us the :
  - Eta squared Gender: 13.24 / 62.29 = 0.21 = 21%
  - Eta squared Employment Status: 19.58 / 62.29 = 0.31 = 31%
- Interpretation:
  - 21 percent of all variance in the dependent variable "happiness" is attributable to gender
  - 31 percent of all variance in the dependent variable "happiness" is attributable to employment status

    → most important main effect

Interpreting effect size

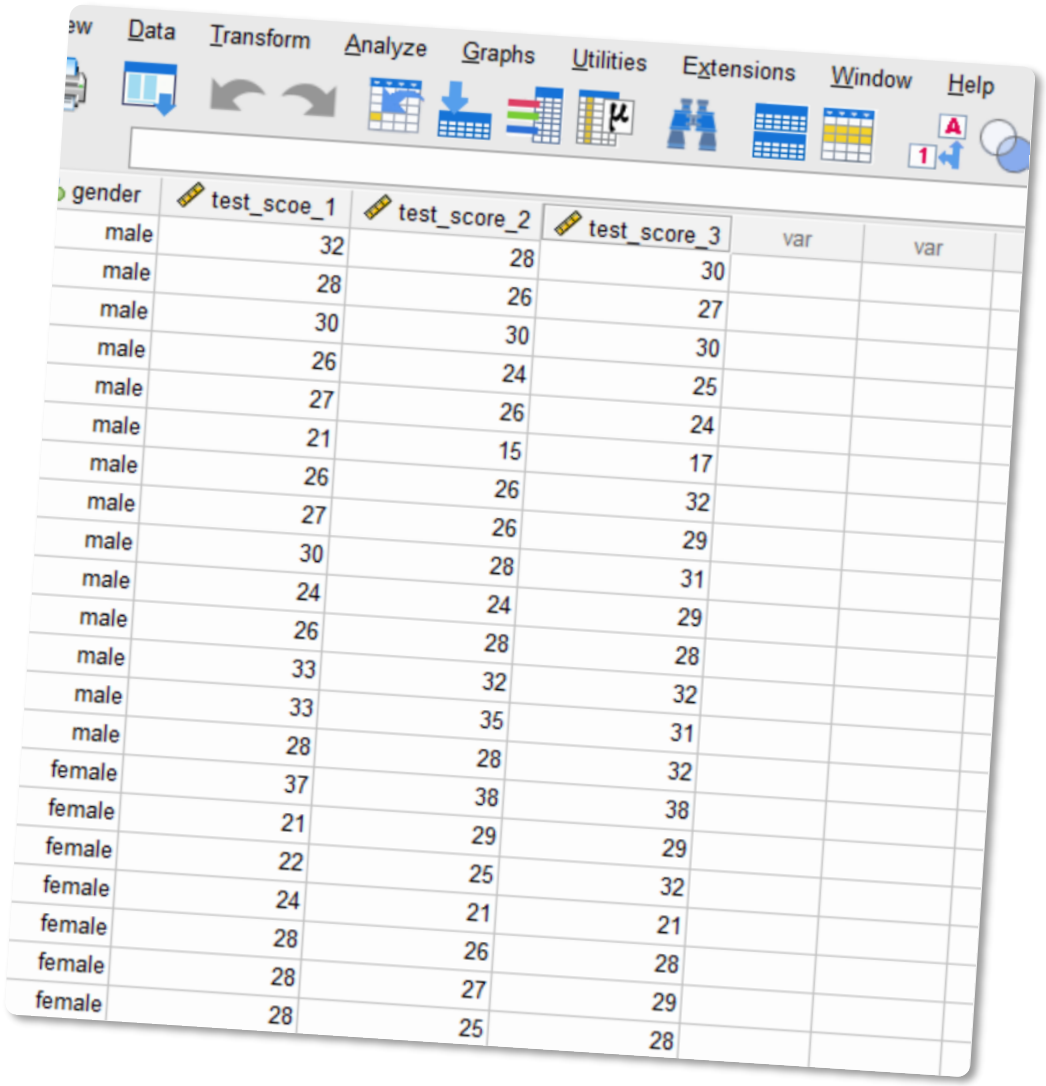
- Cohen's d:
  - d=0.2 (small), d=0.5 (medium), d=0.8 (large)
  - ➡ If two groups don't differ by at least 0.2 standard deviations, the difference of both means is trivial, even if the results are significant
- η^2:
  - η^2=0.01 (small), η^2=0.06 (medium), η^2=0.14 (large)

SPSS

Background:
40 participants in a 2 month vocabulary training program. Participants are tested on their reading comprehension at three different time points (pre, midway & post intervention effects)

Research Question:
Are the test scores different between test 1 and test 3?

Which parametric test is suitable for the above study design?



gender	test_scoe_1	test_score_2	test_score_3	var	var
male	32	28	30		
male	28	26	27		
male	30	30	30		
male	26	24	25		
male	27	26	24		
male	21	15	17		
male	26	26	32		
male	27	26	29		
male	30	28	31		
male	24	24	29		
male	26	28	28		
male	33	32	32		
male	33	35	31		
male	28	28	32		
female	37	38	38		
female	21	29	29		
female	22	25	32		
female	24	21	21		
female	28	26	28		
female	28	27	29		
female	28	25	28		

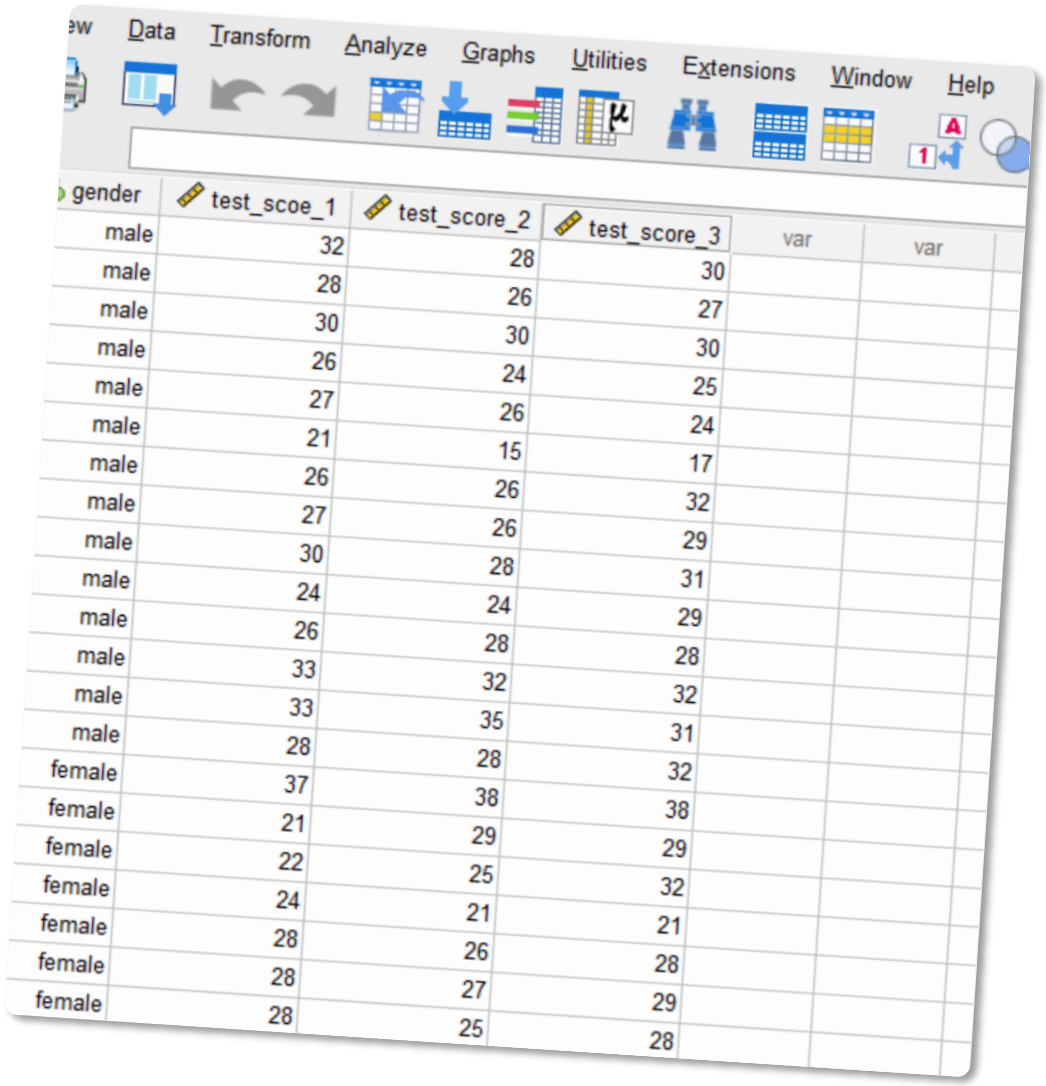
Dependent sample t-test

In SPSS

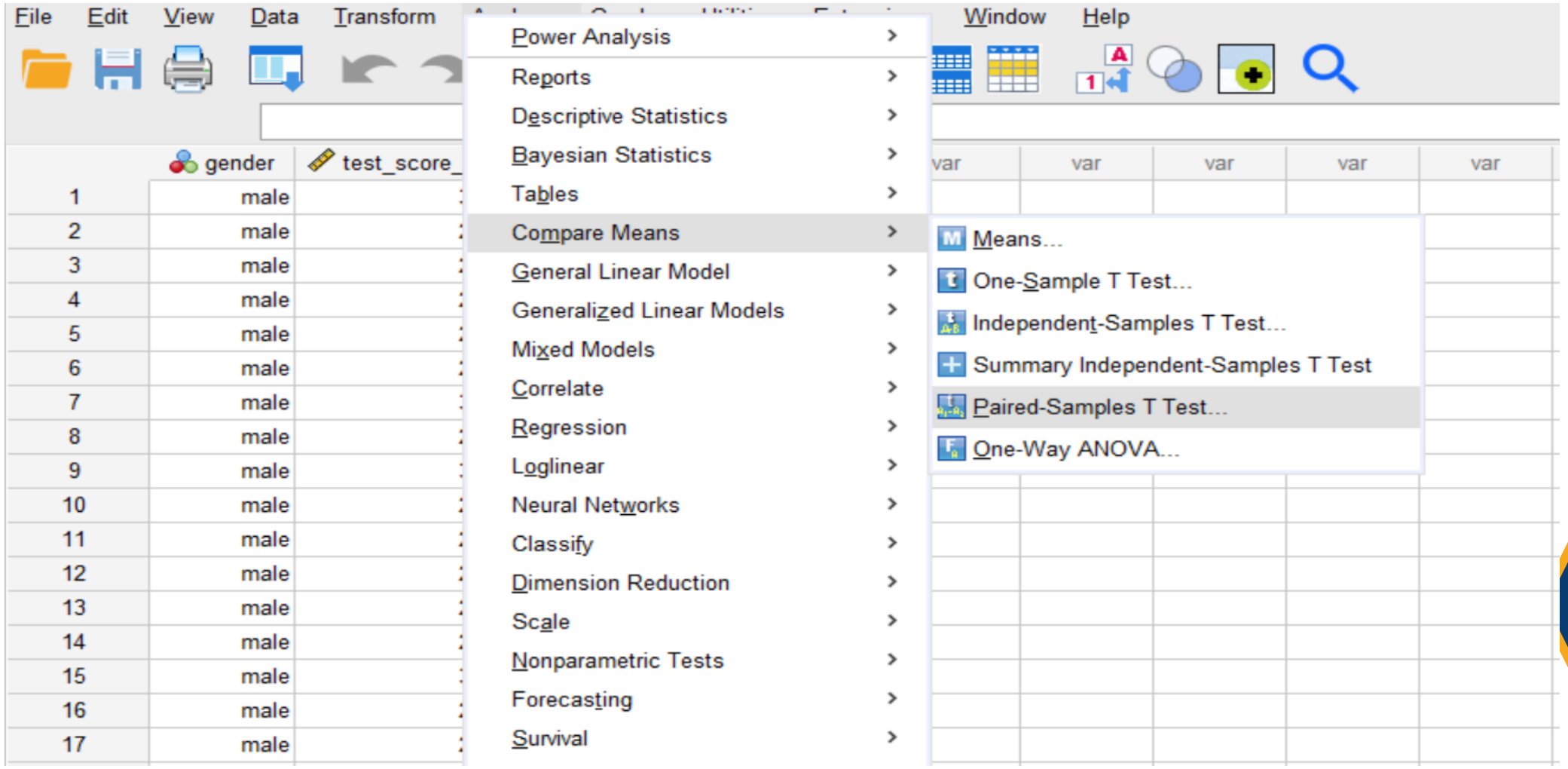
Research Question:
Are the test scores different between test 1 and test 3?

H_A : μ1 ≠ μ2
→ The means of test scores of test 1 and test 3 are not equal and the observed difference is not likely to have occurred by chance alone.

H_0 : μ1 = μ2
→ The means of test scores of test 1 and test 3 are equal.



gender	test_scoe_1	test_score_2	test_score_3	var	var
male	32	28	30		
male	28	26	27		
male	30	30	30		
male	26	24	25		
male	27	26	24		
male	21	15	17		
male	26	26	32		
male	27	26	29		
male	30	28	31		
male	24	24	29		
male	26	28	28		
male	33	32	32		
male	33	35	31		
male	28	28	32		
female	37	38	38		
female	21	29	29		
female	22	25	32		
female	24	21	21		
female	28	26	28		
female	28	27	29		
female	28	25	28		



Analyze → Compare Means → Paired-Samples T Test

gender [gender]
amount of correctly translated word... 1
amount of correctly translated word...
amount of correctly translated word... 2

Paired Variables:

Pair	Variable1	Variable2
1	amount of cor...	amount of co...
2		

Options...
Bootstrap...

Estimate effect sizes

Calculate standardizer using

- (•) Standard deviation of the difference
- ( ) Corrected standard deviation of the difference
- ( ) Average of variances

3 OK Paste Reset Cancel Help

Dependent-samples t-test

SPSS Output

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	amount of correctly translated words in test1	27,53	40	5,208	,824
	amount of correctly translated words in test3	29,18	40	4,924	,779

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	amount of correctly translated words in test1 & amount of correctly translated words in test3	40	,775	,000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	amount of correctly translated words in test1 - amount of correctly translated words in test3	-1,650	3,409	,539	-2,740	-,560	-3,062	39	,004

We can see that there is a statistically significant difference between both test scores. We now want to find out how much the vocabulary training program worked.
We can calculate the effect size ourselves…

$$d = \frac{\mu_{experimental} - \mu_{control}}{\sigma} = \frac{29.18 - 27.53}{3.409} = 0.48$$

Dependent-samples t-test

Effect Size

	Group 1	Group 2
Mean	27,53	29,18
Standard Deviation	5,208	4,924
Correlation	0,775	
Effect Size $d_{Repeated\ Measures}$	0.472	
Effect Size $d_{Repeated\ Measures,\ pooled}$	0.485	
Effect Size $d_{Individual\ Groups}$	0.317	

… or we can use another platform, e.g:
https://www.psychometrica.de/effect_size.html
4. Effect size estimates in repeated measures design

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	amount of correctly translated words in test1 - amount of correctly translated words in test3	-1,650	3,409	,539	-2,740	-,560	-3,062	39	,004

$t(39) = -3.062$, $p = 0.004$; $d = 0.485$

There is a statistically significant difference between both test scores. The vocabulary training program had a medium effect. We can discard the H_0 which stated that the means of the test scores of test 1 and test 3 are equal

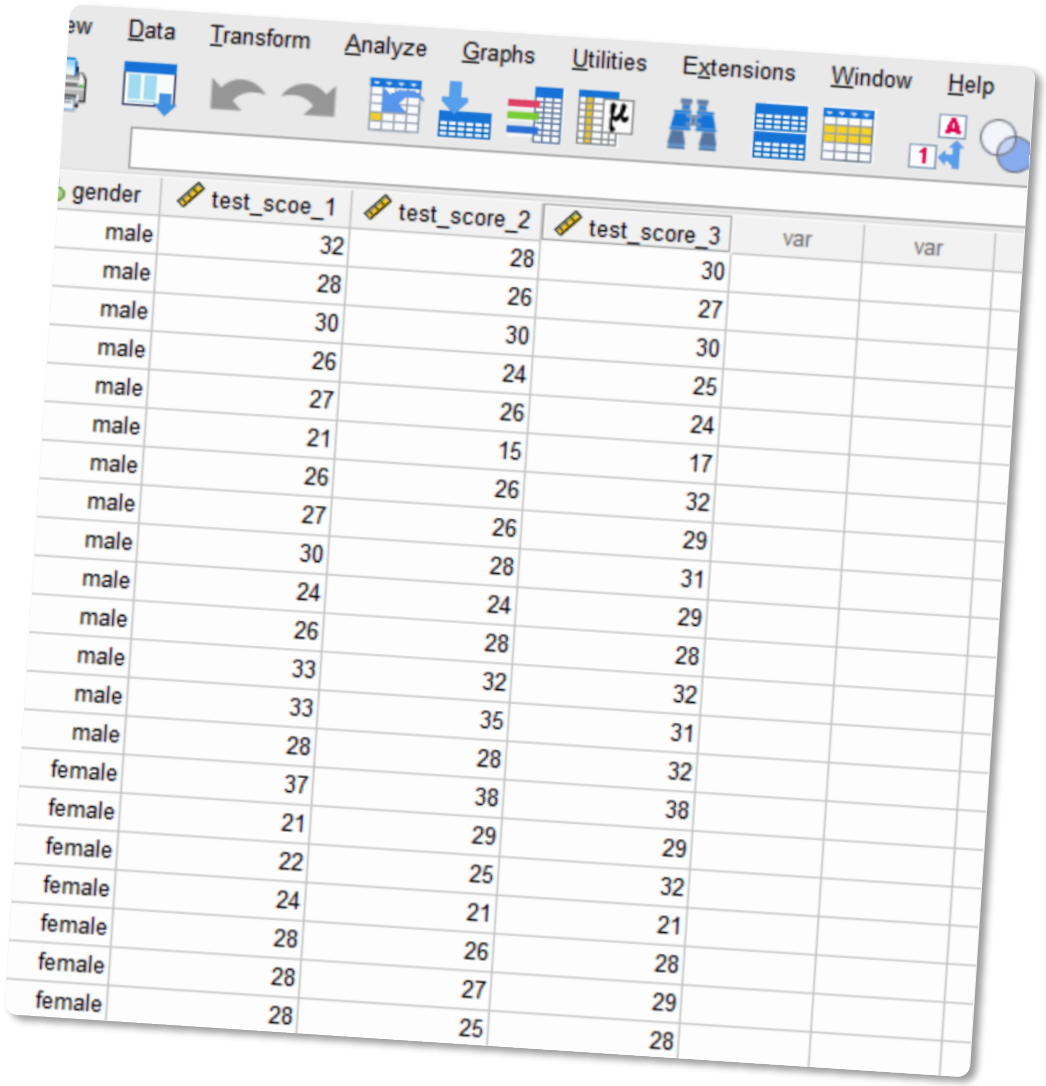
SPSS

READY?

Background:
40 participants (20 male, 20 female) in a 2 month vocabulary training program. Participants are tested on their reading comprehension at three different time points (pre, midway & post intervention effects)

Research Question:
Does the gender of a participant have an effect on the test scores (all three tests)?

Which parametric test is suitable for the above study design?



gender	test_scoe_1	test_score_2	test_score_3	var	var
male	32	28	30		
male	28	26	27		
male	30	30	30		
male	26	24	25		
male	27	26	24		
male	21	15	17		
male	26	26	32		
male	27	26	29		
male	30	28	31		
male	24	24	29		
male	26	28	28		
male	33	32	32		
male	33	35	31		
male	28	28	32		
female	37	38	38		
female	21	29	29		
female	22	25	32		
female	24	21	21		
female	28	26	28		
female	28	27	29		
female	28	25	28		

Within-subjects ANOVA In SPSS

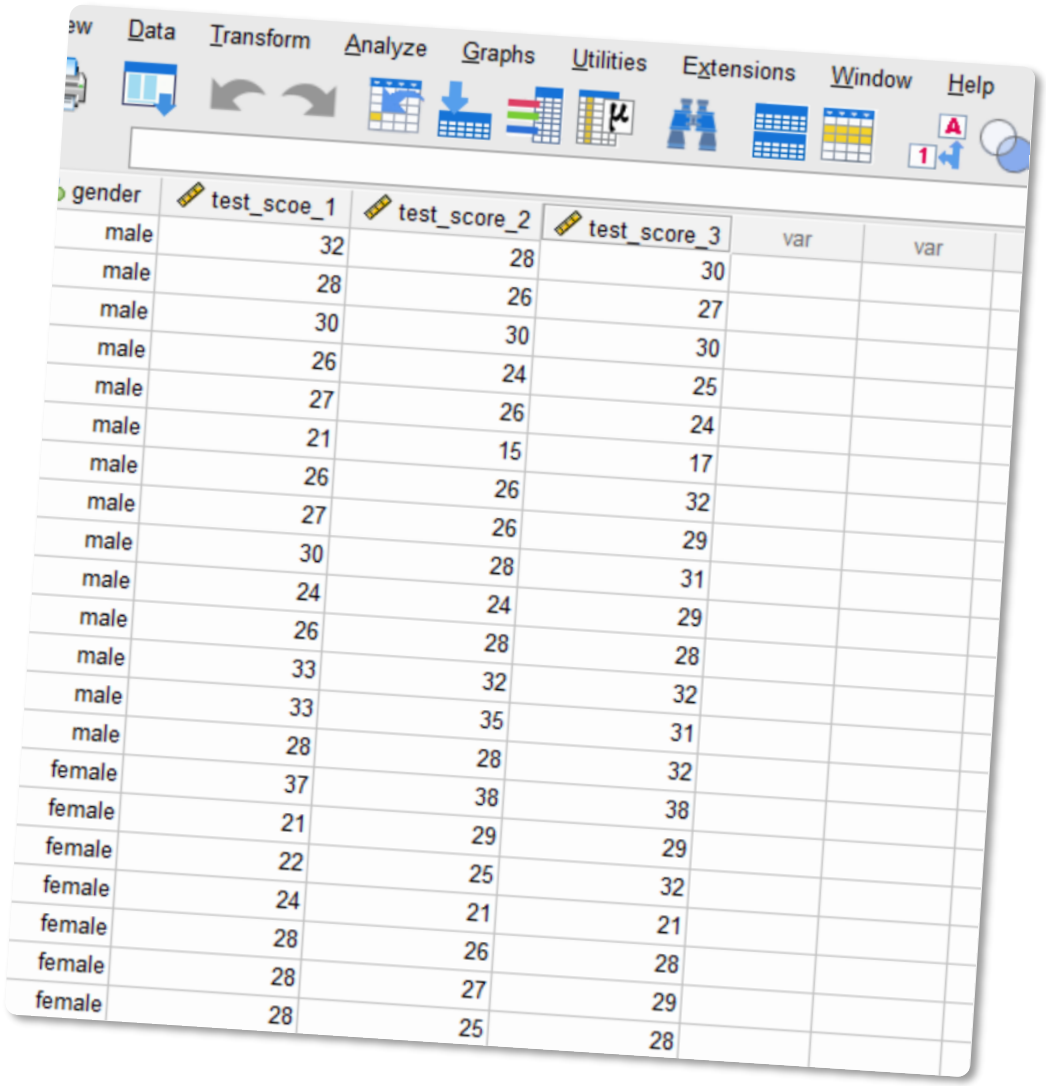
Research Question:
Does the gender of a participant have an effect on the test scores (all three tests)?

H_{A1} : There is an interaction effect between gender and time of testing on the test scores.

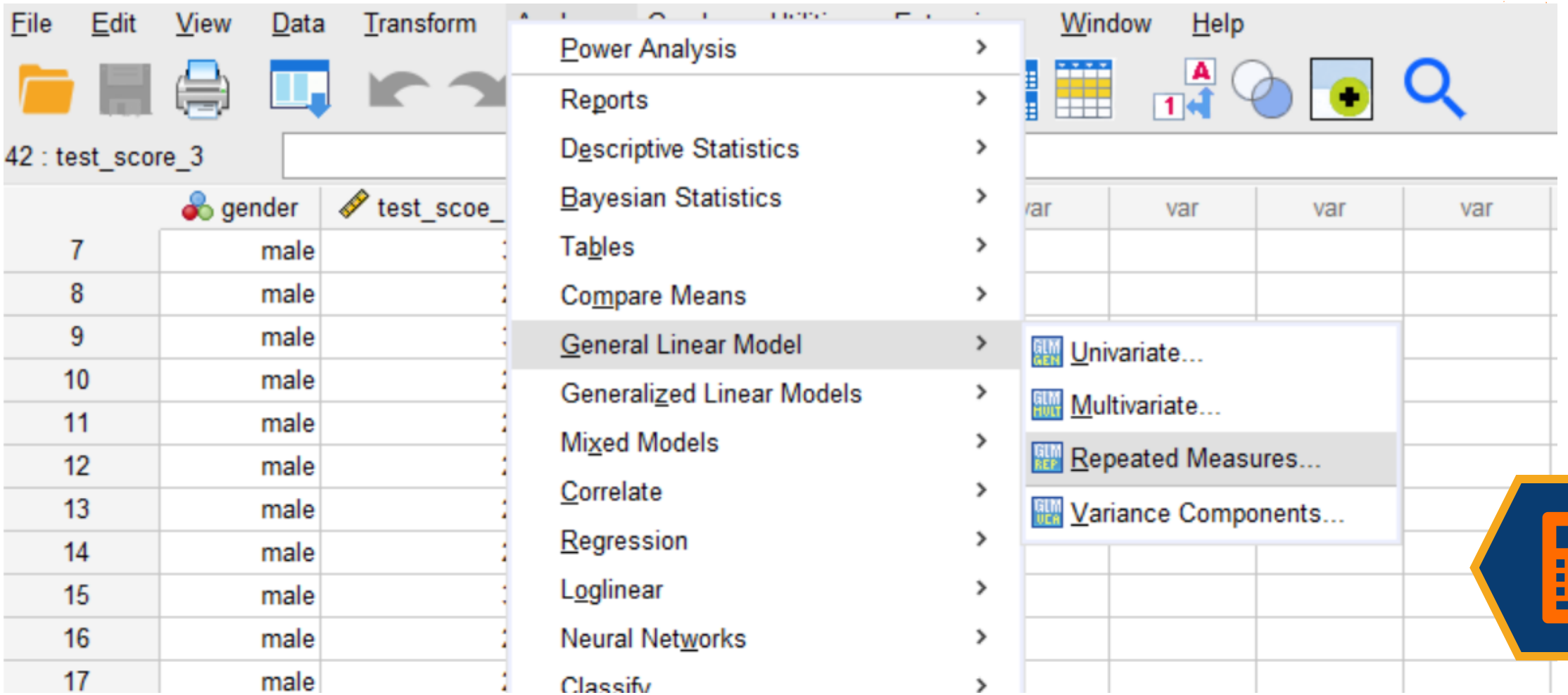
H_{01} : There is no interaction effect between gender and time of testing on the test scores

H_{A2} : There is a main effect of the test time on the test scores.

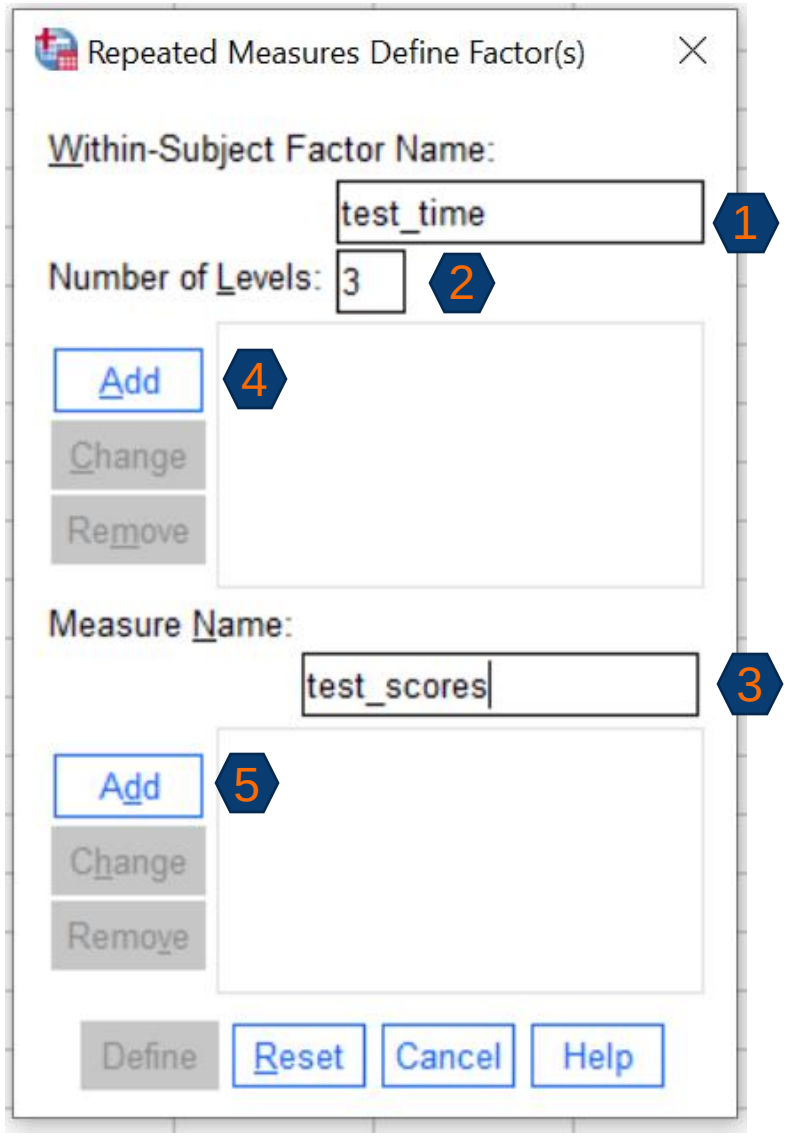
H_{02} : There is no main effect of the test time on the test scores

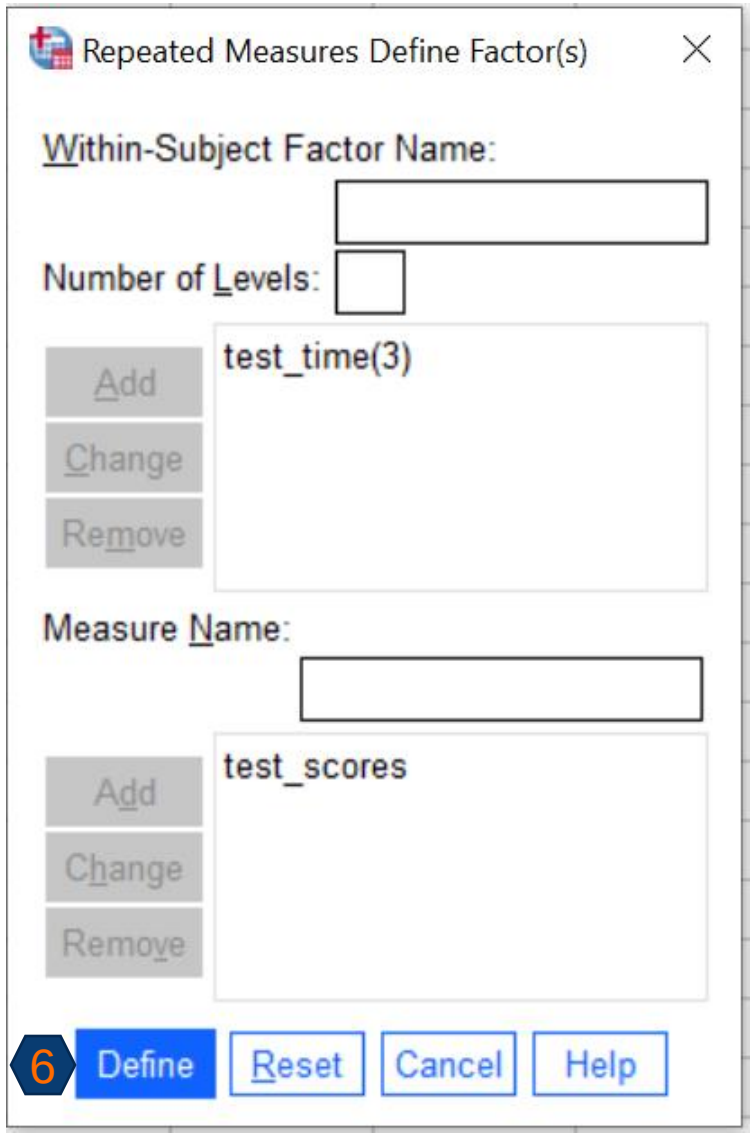


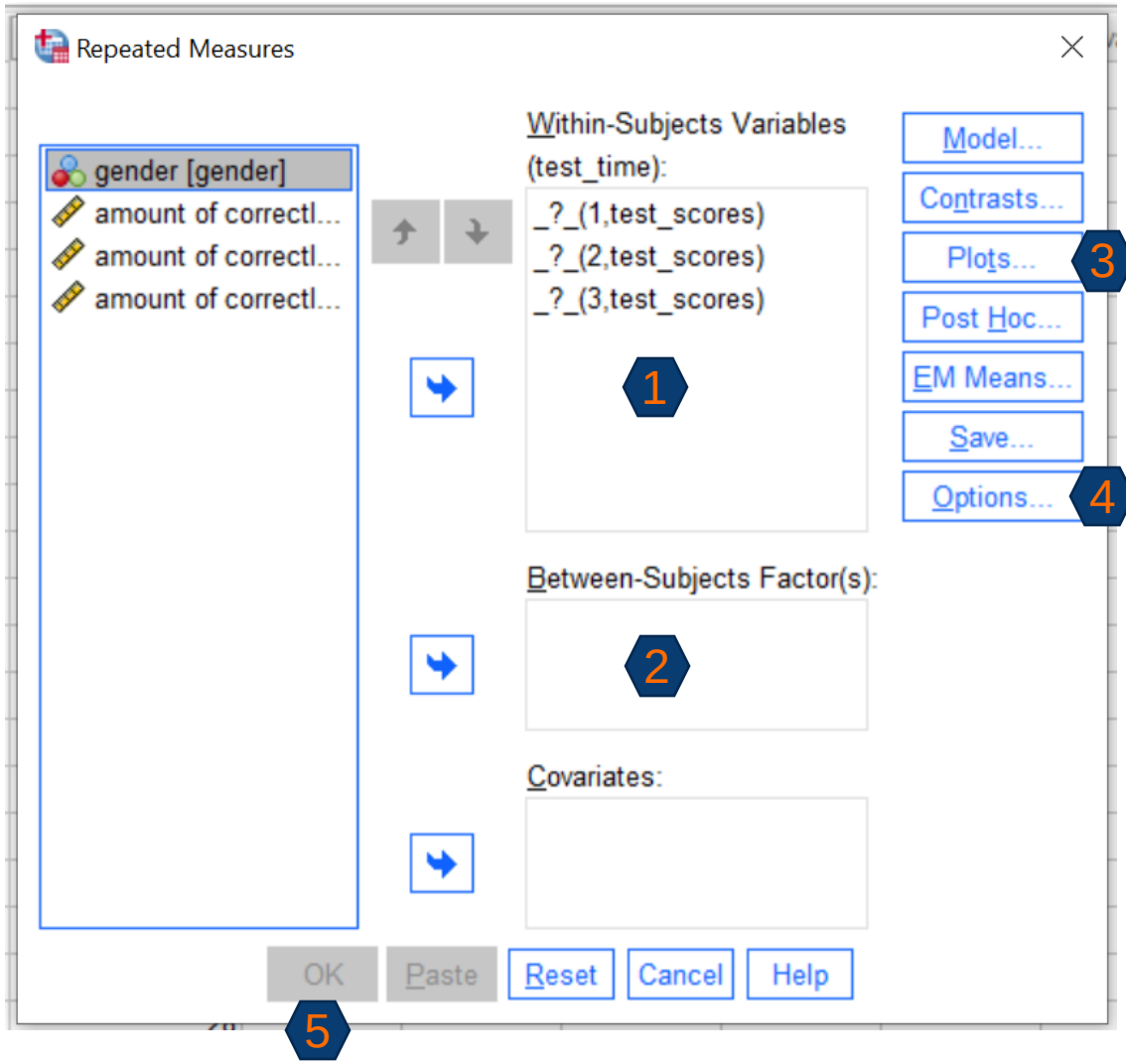
gender	test_scoe_1	test_score_2	test_score_3	var	var
male	32	28	30		
male	28	26	27		
male	30	30	30		
male	26	24	25		
male	27	26	24		
male	21	15	17		
male	26	26	32		
male	27	26	29		
male	30	28	31		
male	24	24	29		
male	26	28	28		
male	33	32	32		
male	33	35	31		
male	28	28	32		
female	37	38	38		
female	21	29	29		
female	22	25	32		
female	24	21	21		
female	28	26	28		
female	28	27	29		
female	28	25	28		

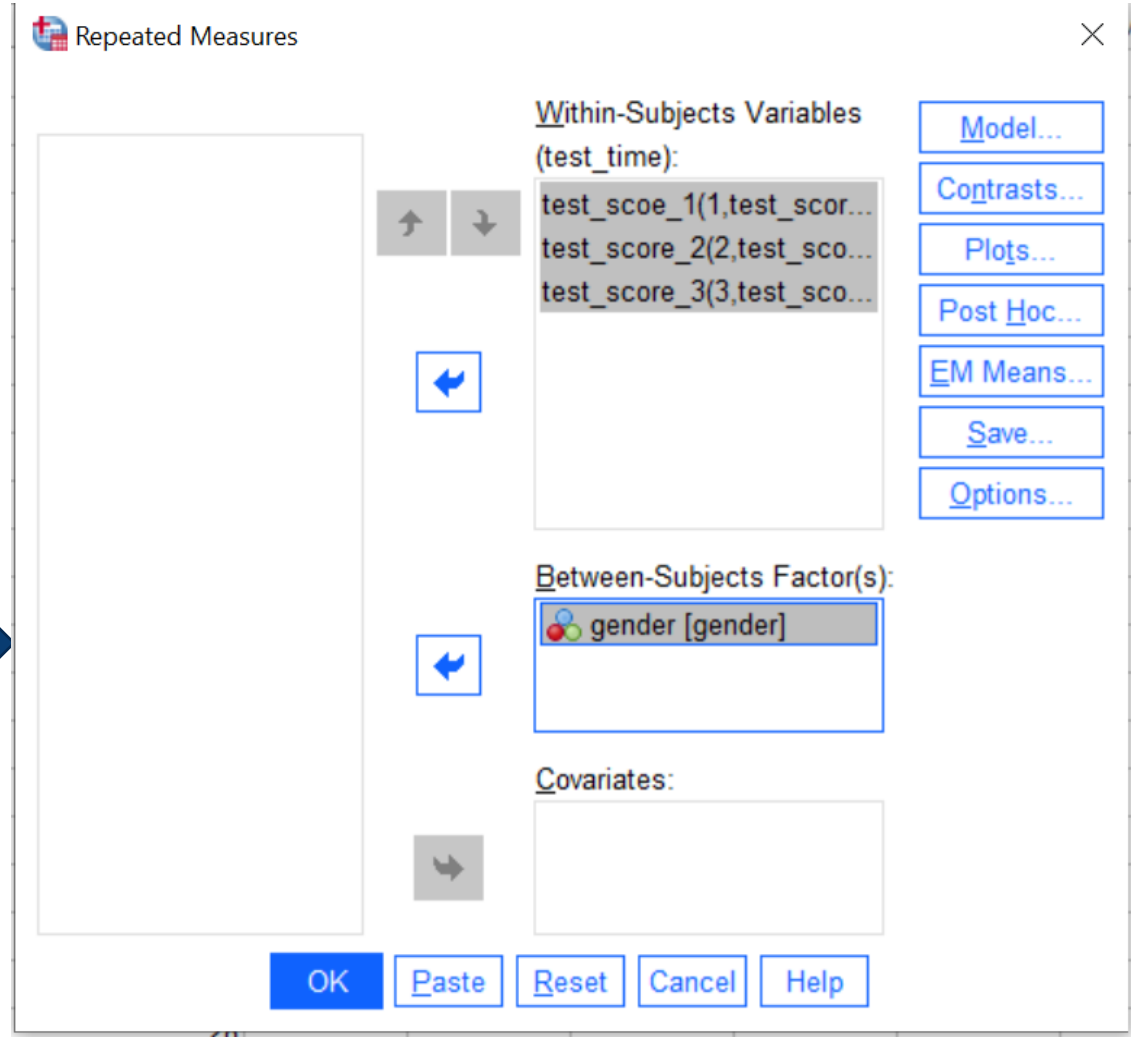


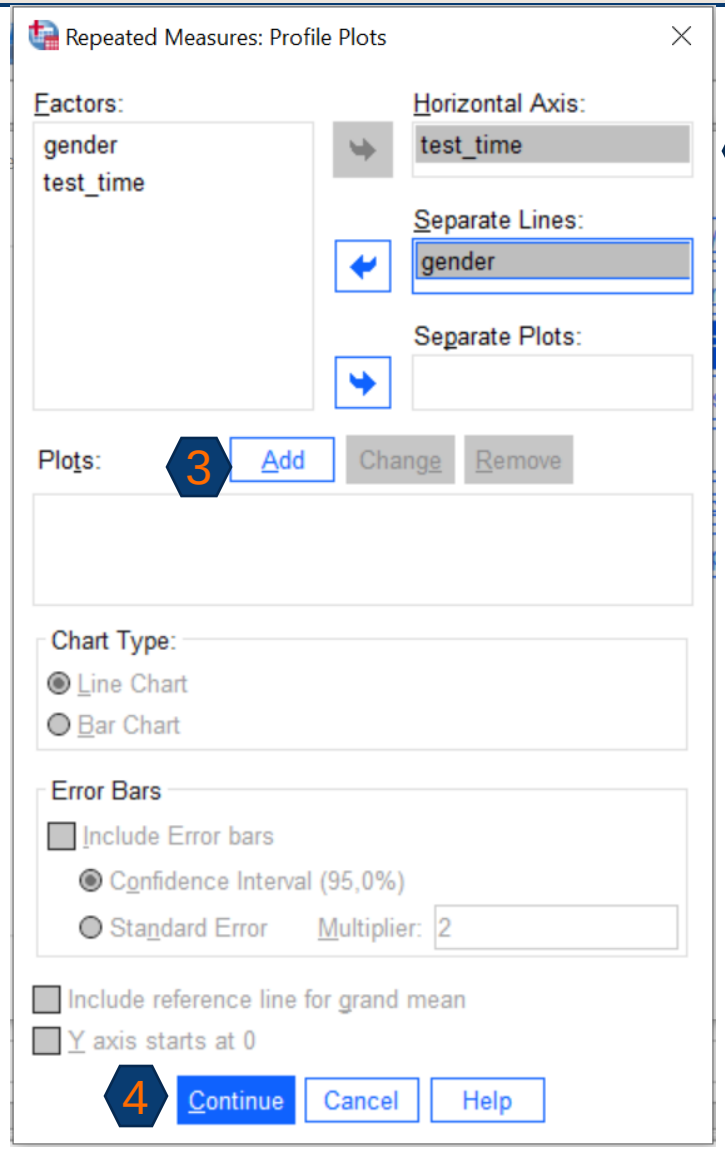
Analyze → General Linear Model → Repeated Measures


Repeated Measures Define Factor(s)
Within-Subject Factor Name:
test_time
1
Number of Levels:
3
2
Add
4
Change
Remove
Measure Name:
test_scores
3
Add
5
Change
Remove
Define
Reset
Cancel
Help

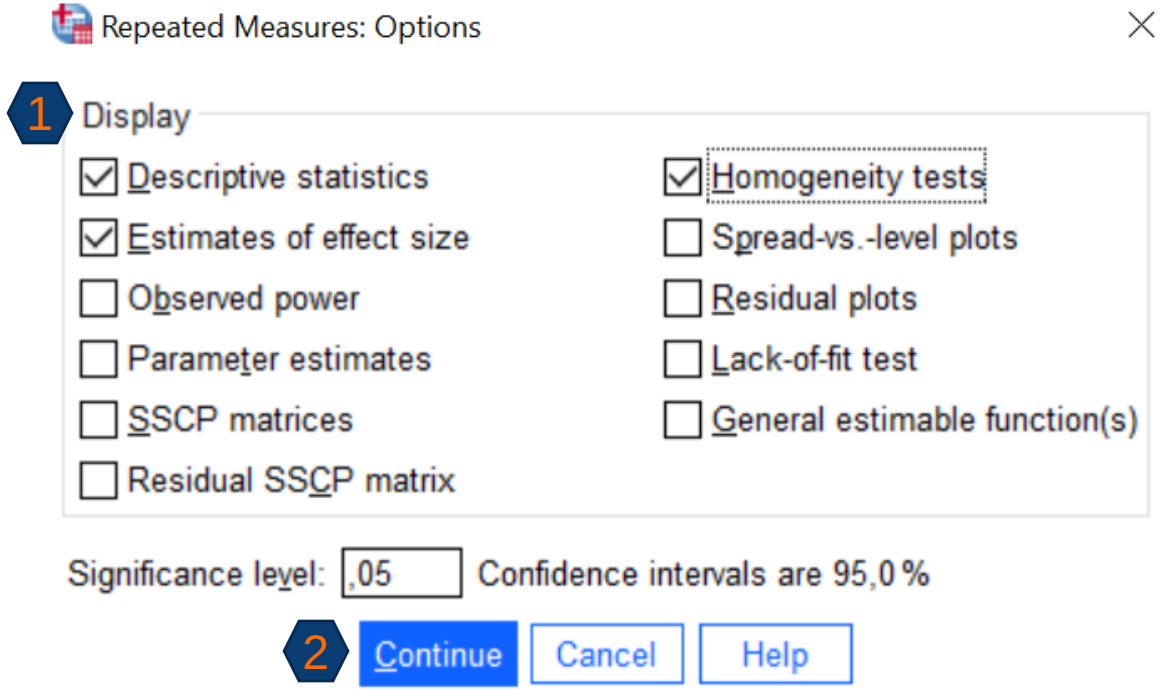

Repeated Measures Define Factor(s)
Within-Subject Factor Name:
Number of Levels:
Add
Change
Remove
test_time(3)
Measure Name:
Add
Change
Remove
test_scores
6
Define
Reset
Cancel
Help


Repeated Measures
gender [gender]
amount of correctl...
amount of correctl...
amount of correctl...
Within-Subjects Variables (test_time):
_?_(1,test_scores)
_?_(2,test_scores)
_?_(3,test_scores)
Model...
Contrasts...
Plots...
Post Hoc...
EM Means...
Save...
Options...
Between-Subjects Factor(s):
Covariates:
OK
Paste
Reset
Cancel
Help
1
2
3
4
5


Repeated Measures
Within-Subjects Variables (test_time):
test_scoe_1(1,test_scor...
test_score_2(2,test_sco...
test_score_3(3,test_sco...
Model...
Contrasts...
Plots...
Post Hoc...
EM Means...
Save...
Options...
Between-Subjects Factor(s):
gender [gender]
Covariates:
OK
Paste
Reset
Cancel
Help


Repeated Measures: Profile Plots
Factors:
gender
test_time
Horizontal Axis:
test_time
1
Separate Lines:
2
gender
Separate Plots:
Plots:
3
Add
Change
Remove
Chart Type:
Line Chart
Bar Chart
Error Bars
Include Error bars
Confidence Interval (95,0%)
Standard Error
Multiplier:
2
Include reference line for grand mean
Y axis starts at 0
4
Continue
Cancel
Help

&


Repeated Measures: Options
1
Display
Descriptive statistics
Estimates of effect size
Observed power
Parameter estimates
SSCP matrices
Residual SSCP matrix
Homogeneity tests
Spread-vs.-level plots
Residual plots
Lack-of-fit test
General estimable function(s)
Significance level:
,05
Confidence intervals are 95,0 %
2
Continue
Cancel
Help

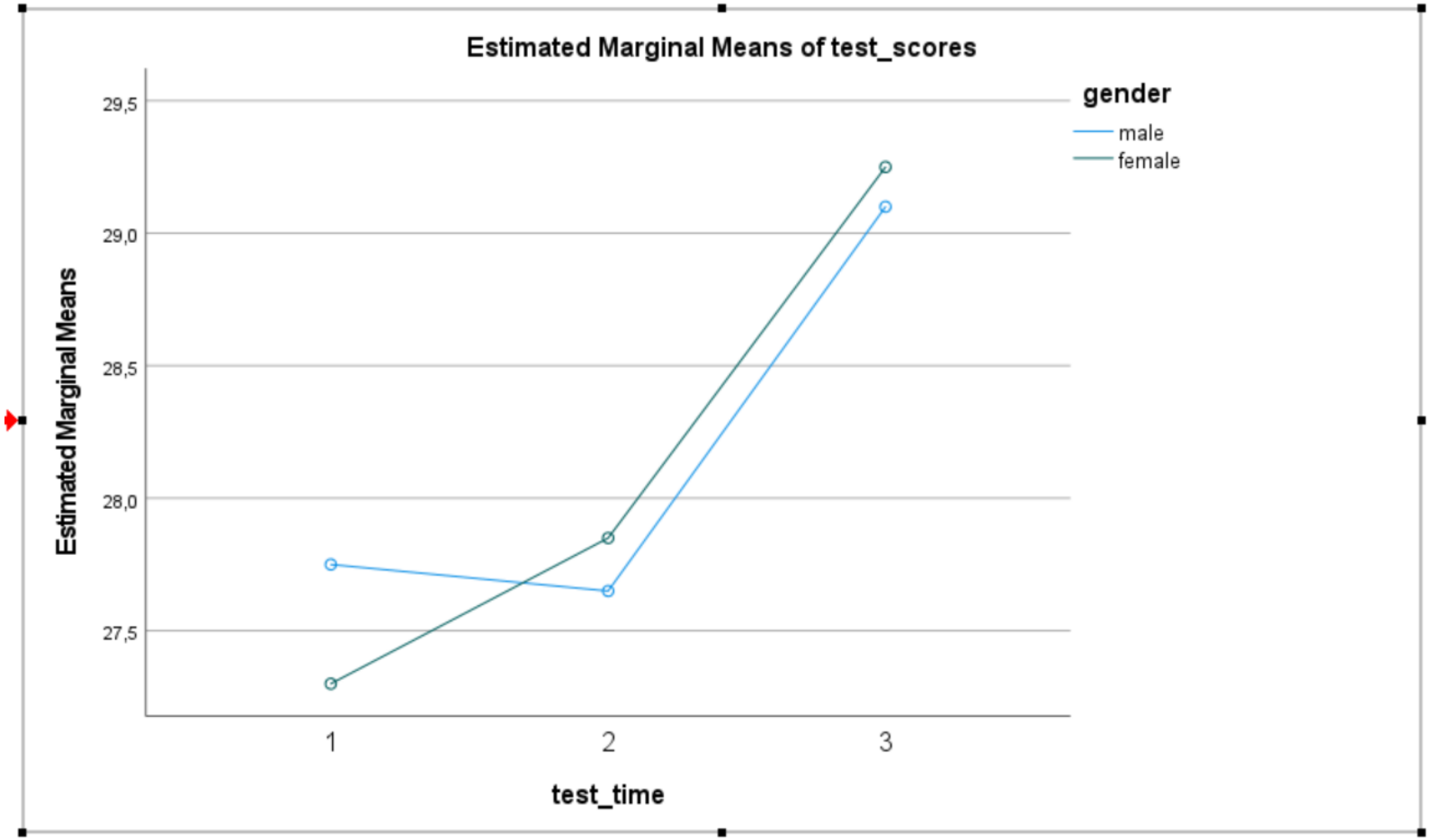
Within-subjects ANOVA

SPSS Output

Descriptive Statistics

	gender	Mean	Std. Deviation	N
amount of correctly translated words in test1	male	27,75	4,141	20
	female	27,30	6,199	20
	Total	27,53	5,208	40
amount of correctly translated words in test2	male	27,65	5,122	20
	female	27,85	5,224	20
	Total	27,75	5,108	40
amount of correctly translated words in test3	male	29,10	4,656	20
	female	29,25	5,300	20
	Total	29,18	4,924	40

Profile Plots



The means are visualized in the profile plots

Output – Mauchly‘s Test of Sphericity

Mauchly's Test of Sphericity[a]

Measure: test_scores

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
test_time	,855	5,810	2	,055	,873	,936	,500

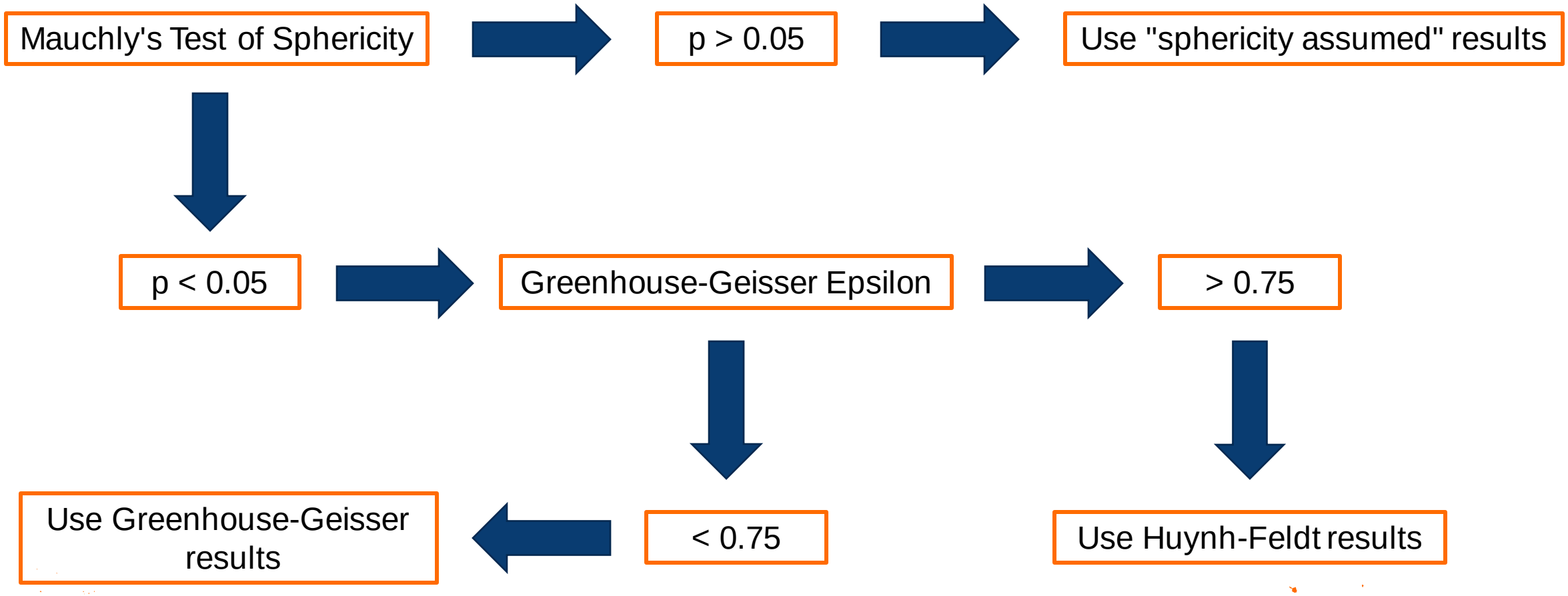
Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept + gender
Within Subjects Design: test_time

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

p > 0.05 → sphericity assumption is met

Mauchly‘s Test of Sphericity
What if the sphericity assumption is not met?



Tests of Within-Subjects Effects

Measure: test_scores

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
test_time	Sphericity Assumed	64,050	2	32,025	6,638	,002	,149
	Greenhouse-Geisser	64,050	1,746	36,679	6,638	,004	,149
	Huynh-Feldt	64,050	1,872	34,223	6,638	,003	,149
	Lower-bound	64,050	1,000	64,050	6,638	,014	,149
test_time * gender	Sphericity Assumed	2,617	2	1,308	,271	,763	,007
	Greenhouse-Geisser	2,617	1,746	1,498	,271	,733	,007
	Huynh-Feldt	2,617	1,872	1,398	,271	,749	,007
	Lower-bound	2,617	1,000	2,617	,271	,606	,007
Error(test_time)	Sphericity Assumed	366,667	76	4,825			
	Greenhouse-Geisser	366,667	66,357	5,526			
	Huynh-Feldt	366,667	71,118	5,156			
	Lower-bound	366,667	38,000	9,649			

$F(2,76)=0.271$, $p=0.763$; $\eta^2=0.007$

There is no interaction between test time and gender. We therefore keep our H_{01} which stated that there is no interaction effect between gender and time of testing on the test scores

Tests of Within-Subjects Effects

Measure: test_scores

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
test_time	Sphericity Assumed	64,050	2	32,025	6,638	,002	,149
	Greenhouse-Geisser	64,050	1,746	36,679	6,638	,004	,149
	Huynh-Feldt	64,050	1,872	34,223	6,638	,003	,149
	Lower-bound	64,050	1,000	64,050	6,638	,014	,149
test_time * gender	Sphericity Assumed	2,617	2	1,308	,271	,763	,007
	Greenhouse-Geisser	2,617	1,746	1,498	,271	,733	,007
	Huynh-Feldt	2,617	1,872	1,398	,271	,749	,007
	Lower-bound	2,617	1,000	2,617	,271	,606	,007
Error(test_time)	Sphericity Assumed	366,667	76	4,825			
	Greenhouse-Geisser	366,667	66,357	5,526			
	Huynh-Feldt	366,667	71,118	5,156			
	Lower-bound	366,667	38,000	9,649			

$F(2,76)=6.638$, $p=0.002$; $\eta^2=0.149$

There is a significant, large main effect of the test time on the test scores. We therefore discard our H_0 which stated that there is no main effect of the test time on the test scores

Let's play Kahoot!

Use your phone and open kahoot.it in your browser

References

- https://www.leeds.ac.uk/educol/documents/00002182.htm
- https://www.psychometrica.de/effect_size.html#transform
- https://www.simplypsychology.org/effect-size.html
- https://www.spss-tutorials.com/spss-paired-samples-t-test/
- https://www.spss-tutorials.com/spss-repeated-measures-anova-example-2/
- https://statistics.laerd.com/statistical-guides/repeated-measures-anova-statistical-guide.php


Thank you for your attention