



Mixed Model Madness

The Mixed Procedure for Cognitive Neuroscience

Fabian van den Berg

MBIC GRADUATE SCHOOL, UNIVERSITY OF MAASTRICHT

This course was set up as part of the MBIC Graduate School at Maastricht University. The text is meant as a support for the lectures, containing all the examples and a little bit more explanation here and there

First Edition, April 2019

Preface

Before getting started I need to get some things out of the way. First and foremost, how nice of you to actually read this, good for you! That aside, the real first thing is that I am no statistician, didn't have the education nor training a real statistician would and should need. What I do have is a very specific set of skills, a set of skills that includes Googling, and I've used this skill-set to learn a lot about Mixed Models. It took me several years to get everything together and I'm hoping this course saves you a lot of the trouble I had.

Mixed Models isn't completely mainstream yet, it's getting there though. Because of that there's not a lot of support for them, no standard procedures, comprehensive guides, or tutorials. The things I cover in this course is stuff I had to figure out on my own, learn how to interpret the output, draw conclusions, write down results, include contrasts, polynomials, and deal with weird data.

A second important point is that everything in here is simplified. I did my best to keep out the math side of the models (honestly, I don't totally comprehend all the math either). I decided to make it a practical course on how to apply Mixed Models to your own data. There are a lot of things that decreased in accuracy in favor of understandability, so technically it's not totally true what I'm saying but it's close enough to not confuse and help understand (plus it helps making the pictures). The purpose is to get the point across, since none of you are statisticians either and probably also don't care about the math.

I would recommend looking into the math if you have the time and motivation, it really helps understand exactly how it all works and justify your choices (then you won't have to rely on others saying what's what).

Lastly, we're using SPSS (I know, I know, we all want to use R). The reason for this is that I can be certain we all know SPSS. It saves the trouble of getting everyone up to speed with R, that's a whole different course. Mixed Models in R is also a bit of a bitch, you'll be a lot more prepared to use them in R once you understand them conceptually. R requires the formula-notation of mixed models, adding fixed and random effects manually. SPSS does this for you and saves you a headache. I'm hoping the understanding will translate to other programs.


The **goal** of the course is not to teach you the details of mixed models, the goal is to give you a working understanding of the method. Knowing some of the tricks will hopefully help you run these models with some confidence, without having to ask for help, and without making silly mistake.



1	Repeated Measures ANOVA	7
1.1	ANOVA Recap	8
1.2	Repeated Measures Recap	9
1.2.1	Within Subject Effect	9
1.2.2	Separate Error Terms	12
1.2.3	Effect of adding Person as random Variable	15
1.2.4	Sphericity Assumption	18
1.3	Repeated Measures ANOVA	20
1.3.1	Multivariate Test	20
1.3.2	Univariate Tests	21
1.3.3	Pairwise Comparisons	22
2	Mixed Models Procedure	23
2.1	RM ANOVA versus Mixed Models	25
2.1.1	The Perfect Design	25
2.1.2	Clustering or Multiple Levels	25
2.1.3	The Case of the Missing Data	26
2.1.4	The Factor of Time	26
2.1.5	Winner: Mixed Models	27
2.2	Example Data	27
2.3	Using Repeated Measures ANOVA: Output	28
2.4	Using the Mixed Procedure: Output	30

3	Building a Model	33
3.1	Covariance Matrices	34
3.2	Covariance Matrices in Mixed Models	37
3.2.1	Compound Symmetry (Univariate method)	37
3.2.2	First Order Autoregressive (AR1)	37
3.2.3	Toeplitz (TP)	38
3.2.4	Unstructured (Multivariate Method)	38
3.3	Which Structure to use?	39
3.3.1	When Unstructured Doesn't work	39
3.3.2	When Unstructured Does work	40
3.4	Assumptions of Mixed Models	40
3.4.1	Linearity	40
3.4.2	Homoscedasticity	41
3.4.3	Multicollinearity	42
3.4.4	Outliers and Normality	42
3.5	Building a Model	44
3.5.1	Comparing two Models	44
3.5.2	When is a model a better fit?	45
3.5.3	Step 0: The Data	46
3.5.4	Step 1: Factor Selection	46
3.5.5	Step 2: Covariance structure selection	47
3.5.6	Step 3: Model Reduction	48
3.5.7	Step 4: The Final Model	49
3.5.8	Step 5: Writing it Down	49
4	Extending the Model	51
4.1	Post-Hoc Testing (Estimated Marginal Means)	53
4.1.1	What are EMMs?	53
4.2	Adding Covariates to the Model	55
4.2.1	Without Covariate	56
4.2.2	With Covariate	59
4.2.3	Dealing with Moderation	62
4.3	Random Effect Models	64
4.3.1	Model 1: Fixed Effects Model	65
4.3.2	Model 2: Random Slope Model	67
4.3.3	Model 3: Random Intercept Model	68
4.3.4	Model 4: Random Intercept and Slope Model	69
4.3.5	Real Example: Random Session Effects	71
5	Customizing the Analysis	73
5.1	Custom Contrasts	74
5.1.1	Custom Contrasts in a GLM	75

5.1.2	Custom Contrasts in the Mixed Procedure	79
5.2	Generalized Estimation Equations (GEE)	88
5.2.1	Running a GEE in SPSS	88
5.2.2	Link Functions	91
6	Appendices	93
6.1	Appendix A: Polynomial Contrast Coefficients	94
6.2	Appendix B: The Mixed Regression Procedure	95

A wide-angle photograph of a city at sunset. The sky is filled with dramatic, colorful clouds in shades of orange, pink, and purple, with the sun low on the horizon. A dark, semi-transparent rectangular box with rounded corners is positioned in the lower-middle part of the image, containing the title text.

1. Repeated Measures ANOVA

Introduction

To begin this course we will first focus on the **Logic of Repeated Measures**. This is a refresher on ANOVA and RM ANOVA. We will look at the effect of adding **Person** as a factor into a model, how taking into account that the same people do the same test multiple times can improve the power of your tests.

We'll also cover the assumption that comes with repeated measures: **Sphericity**. What it means and why it's important. The end of this section will quickly cover the output of the **Repeated Measures ANOVA** in SPSS.

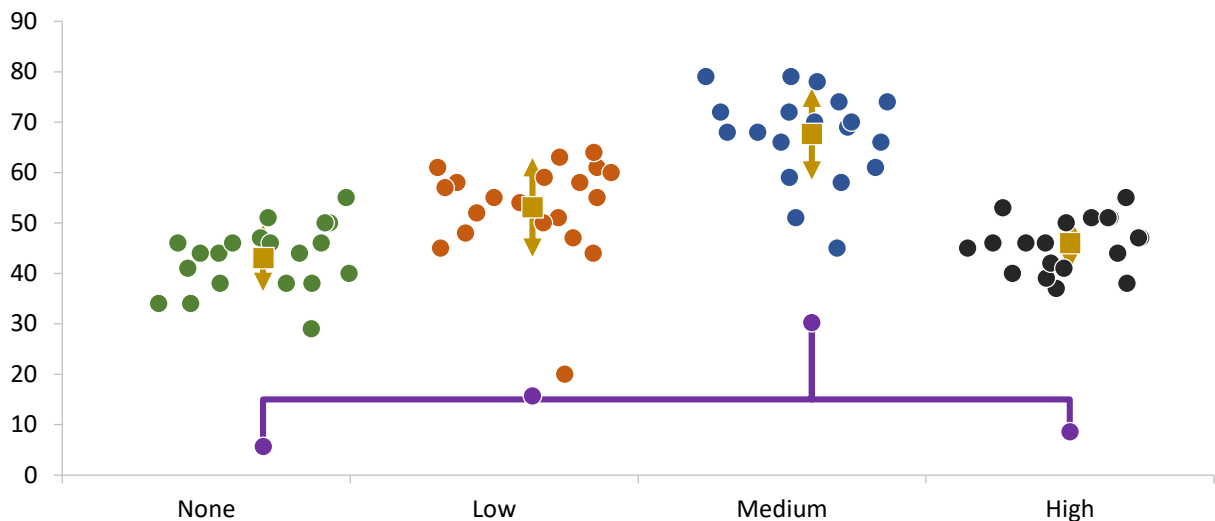
1.1 ANOVA Recap

First a little recap of the normal ANOVA, where we compare groups. We calculate the variance within the groups, which expresses error or anything that isn't explained by what we control for. We also calculate the variance between the groups, which expresses the effect caused by us making the groups. Since we control for this, we make the groups and separate people accordingly, we've biased them. If we look at group 1 we see the green dots representing the people in group 1. Below in purple we see the mean of this group, the golden arrow represents the variation (the standard deviation). Let's ask ourselves why is this purple dot in group 1 lower than the overall average of all the people in our sample (as shown in the line in the bottom)?

- 1) The first reason is because people in group 1 score lower than average. That's our effect and is due to us making the groups. This is what we want to see.
- 2) The second reason is error. If we would take a new group of people we'd get a different average. The average is affected by error, pulling it up or down.

We can ask the same question within the **Group**. Why is this specific person higher or lower than the average of the **Group**?

- 1) Only one reason here, because of error. This person scores higher/lower than the group mean because of factors other than being in the **Group**.



With those two points we can say that **we expect the variance of the groups (MS Between or MS Groups) to express both a **Group** effect and **error****. Whereas **the variance of the error (MS Within or MS Error) is expected to only show the **error****. That is where $\frac{MS(Groups) + MS(Error)}{MS(Error)}$ comes from, and why F is expected to be almost 1 when there is no **Group** effect. Without the **Group** affecting the average we expect the mean of the groups to also express **error**.

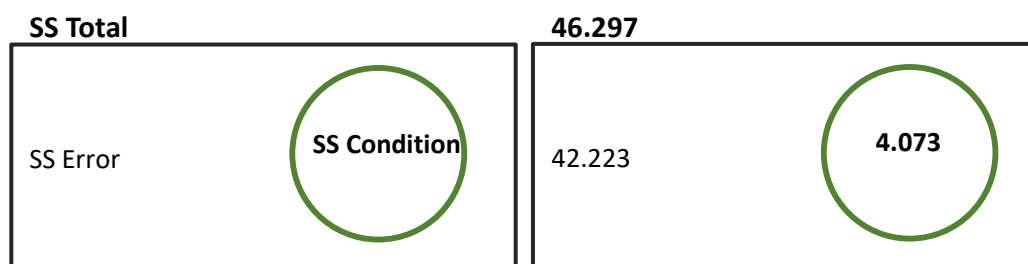
1.2 Repeated Measures Recap

1.2.1 Within Subject Effect

Your computer has no idea what your design or your intention is, it just does the math for you. We'll first analyze some data as if it was a regular ANOVA (this is data and results from Stat-2 which I shameless stole for this example). The way the data was set up is in a univariate format, as you can see in the table below. For you it might be obvious that this shows data from the same participant on the same test at three time-points. But the computer isn't that clever, it doesn't know what you mean with the variable names.

PP	Value	Condition
1	46	1
1	55	2
1	68	3
2	44	1
2	44	2
2	61	3
3	72	1

If we have 40 participants doing the test three times, we have three values per person, but still only 40 participants. In the first analysis the data is analyzed using a normal ANOVA. SPSS will see the data as 120 participants (observations), ignoring the fact that scores come from the same person. It averages and compares the data from condition 1, 2, and 3. Sounds about right, doesn't it? That's what we want to do, compare the three conditions. It results in the image below, condition explaining some variation in our data, the rest is error. The ANOVA also tells us that this effect is significant, wonderful! But we are forgetting something, these conditions are not independent. Condition 1, 2, and 3 are correlated because they come from the same people.



Tests of Between-Subjects Effects

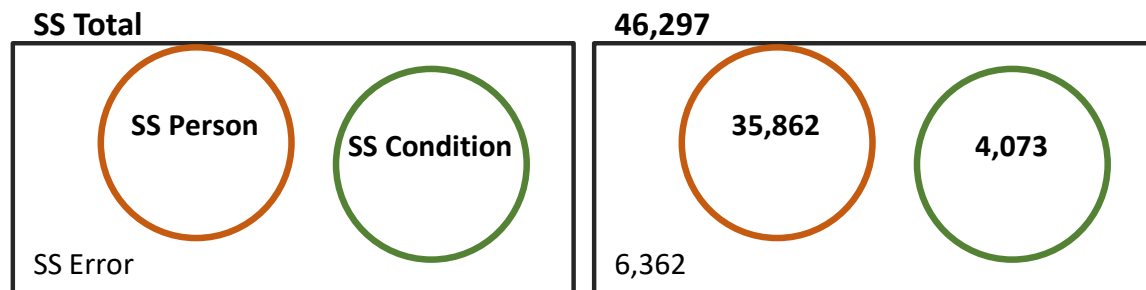
Dependent Variable: RTMEAN

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4.073 ^a	2	2.037	5.643	.005
Intercept	1741.289	1	1741.289	4825.065	.000
CONDITION	4.073	2	2.037	5.643	.005
Error	42.223	117	.361		
Total	1787.586	120			
Corrected Total	46.297	119			

a. R Squared = .088 (Adjusted R Squared = .072)

People can be very consistent, even if some time has passed. Part of our error there is explained by the fact that we have the same **Person** doing the same test three times. Someone who is inherently faster than other people will always have faster reaction times in all three tests. There's a correlation between conditions caused by a **Person** effect. This variation caused by people is ignored in the ANOVA. We have a Within Subject Design, so we should treat it as such. So, let's add **Person** to the model shall we? **Person** would be a random variable and we can add that to the UNIANOVA as such. Why is **Person** a random variable? Because the population of people is far greater than the sample we have and we are interested in the general **Person** effect (see Box 1).

Looking at the next output shows us the results when we include **Person** as random variable. It should be very clear that the effects have changed. The first thing you should notice is that **Person** can explain a lot of variation in the data, explaining a large chunk of the error. So the inclusion of **Person** already lowers SSE, which we like. The second thing is that the model now recognizes that each **Person** has three records; we only have 40 people. The Degrees of Freedom have undergone a change, from using 120 independent records to using 40 with three dependent values each.



Tests of Between-Subjects Effects

Dependent Variable: Value

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	1741.289	1	1741.289	1893.681	.000
	Error	35.862	39	.920 ^a		
CONDITION	Hypothesis	4.073	2	2.037	24.969	.000
	Error	6.362	78	.082 ^b		
Person	Hypothesis	35.862	39	.920	11.274	.000
	Error	6.362	78	.082 ^b		

a. MS(PP) | b. MS(Error)

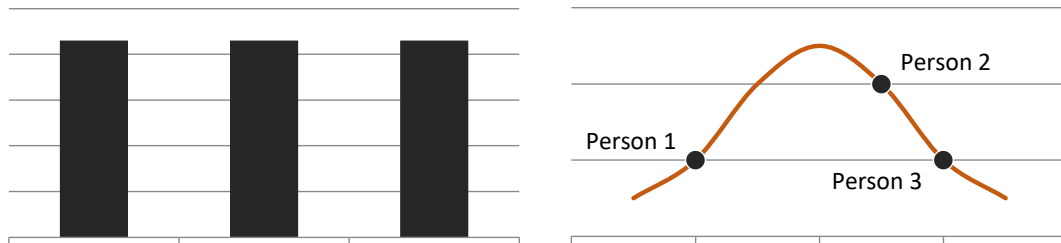
Box 1. Fixed and Random Factors

A **Fixed Factor**, as the name implies, is **fixed** in the sample and the population we are interested in. If we make three groups then that's it, there are only those three groups and we can only conclude something about those three groups. Fixed variables capture all possible values that this variable can take. A **Random factor** is only a sample of the possible values that this variable can take.

If we test the **fixed effect** of medicine we can make three groups: *3gr* – *6gr* – *8gr*. We compare the groups and conclude that the higher the dosage the better the result is. If we treat dosage as a **fixed factor** we can only conclude that 8gr is better than 6 or 3 grams, but nothing about the values in between.

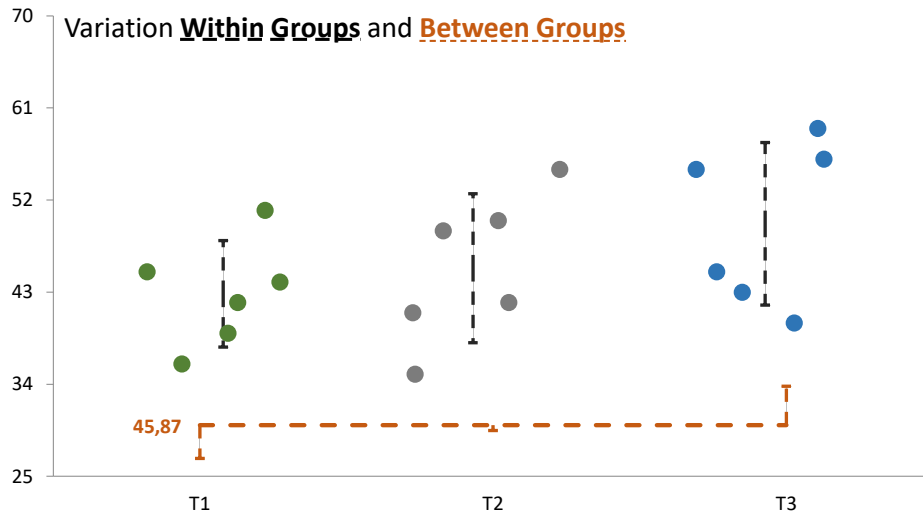
If we add Dosage as a **random variable** we assume that our variable is a subset, a sample, of all the possible dosages in the population. Then we can say that higher is better, rather than limiting ourselves to our 3 fixed groups. In this case we don't assume that there's an effect of 3, 6, and 8gr but that there's a general medicine effect and our three groups all represent this general effect in different ways.

It's the same for **Person**. We can add **Person** as a **fixed effect** and then we know whether person 1 differs from the other people in the data, which they most likely will. If instead we add **Person** as a **random factor** we see the people in our data as a sample from a larger population of people and we get an estimate of the general **Person** effect.

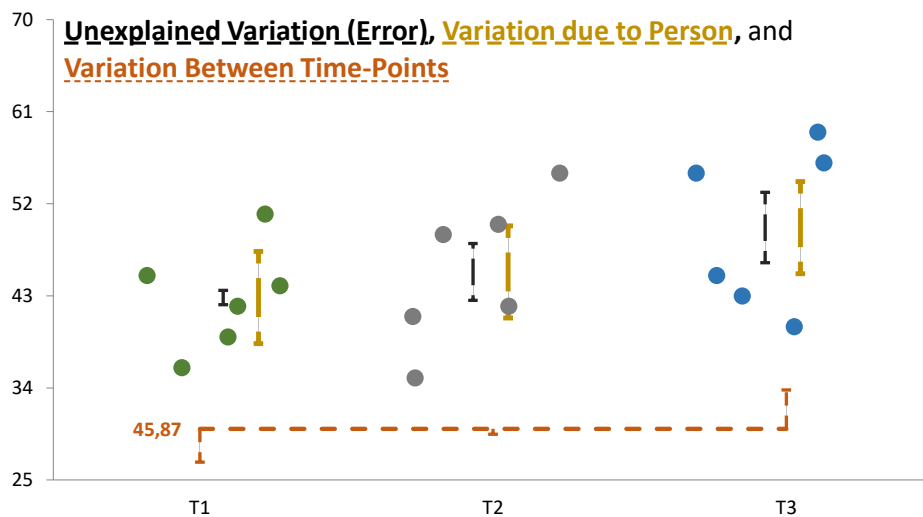


1.2.2 Separate Error Terms

Then to address the elephant in the room, we have separate **Error** terms for our effects. That's because our calculation has changed a bit. Let's think back to the One-Way ANOVA model. There we had two things affecting the scores, their group and other factors such as individual differences (or things like the weather). We expected MSG to be **Group + Error** and we expected MSE to be just **Error**. We expected $F = \frac{MS_{Groups} + MS_{Error}}{MS_{Error}}$ which would lead to approximately 1 if there was no **Group** effect and larger than 1 with a **Group** effect.



But now our model has something extra. The groups are no longer independent; we have the same people in each group. To account for this, we added an extra **random** factor of **Person**. Part of our variation can now be explained by the **Person** factor. This variation is also the same in each group, each group has the same people in them and the model assumes that this variation doesn't change (the differences between the Mean and Person 1 is the same at T1, T2, and T3).



We still want to know if there is a difference between time-points, and this difference has to be larger than differences caused by the people we used. We ask ourselves again, why is the average of Time-point 1 lower than the overall average? But this time our reasons become a bit more specific.

- 1) The first reason is because people at Time-Point 1 scored lower than average. That's our effect and is due to being in Time-Point 1. This is what we want to see.
- 2) The second reason is **Error**. There's a lot of stuff we're not controlling for in our experiment. Perhaps people didn't sleep so well before, perhaps it was a depressing day, maybe it was Friday and people weren't paying attention. We don't know, we can't explain it.
- 3) The third reason is new. This is because these specific people are in Time-Point 1. We've added a **Person** effect, and you being you affects your score. Some people simply score better or worse for some reason, and they'll do this consistently.

We also ask again, why does this specific person have a lower/higher score than the average within Time-point 1?

- 1) First we have **Error**; we don't know what this person did before. It might just not have been his or her day or they performed better than usual because they slept great that night. We don't know for sure.
- 2) The second reason is also the **Person** effect, it's because it is this specific individual in Time-Point 1. This specific person could be generally better at the task, consistently score above average.

We can say that we **expect the variance of the groups (MS Between or MS Groups) to express a Time-Point Effect, an effect of these specific people at their specific Time-Points, and error. The variance of the error (MS Within or MS Error) is expected to express the variance caused by these specific people in this specific time-point, and error.**

Now we have a slightly different F-ratio calculation:

$$F = \frac{MS(Time) + MS(Time * PP) + MS(Error)}{MS(Time * PP) + MS(Error)}$$

If there is no **Group** effect then F will still be 1 because we divide the same numbers (MS[Time*Person] will disappear because it cannot be separated from MS[Error]; see Box 2.).

Sure, you could ignore the **Person** effect, but then your estimate of the effect would be off. To be significant the effect of **Group** needs to be larger than the effect of **Person** and **Error**, not **just Error**. If we don't take this into account then the result might just as well be due to the **Person** effect.

The **Person** effect can be computed using the regular **Error** term. Here we want to know if the differences between people is more than the differences we have unexplained (like weather, having a bad day, or just random chance). The difference between **Group** or time-points doesn't matter here, we assume that this effect is the same in each **Group** (and the average of the **Person** variation in each group will be the same as a single group). Calculating the effect of **Person** is possible, but it doesn't make a lot of sense. You don't really care about differences between participants, since those change with each sample (random variable after all).

$$F_{Person} = \frac{MS_{PP} + MS_{Error}}{MS_{Error}} \quad F_{Cond*Person} = \frac{MS_{PP*Group} + MS_{Error}}{MS_{Error}}$$

To get our F-value we divide the Mean Square of that term with the Mean Square of that term's interaction with **Person**. We compare the effect between **time-points** to the effect of **time-point*Person**. You'd expect that **time** causes more variation than the fact we use a random set of people.

$$F_{condition} = \frac{MS_{Condition}}{MS_{Condition*PP}} \quad F_{Person} = \frac{MS_{Person}}{MS_{Error}} \quad F_{Cond*Person} = \frac{MS_{PP*Group}}{MS_{Error}}$$

But where is the interaction? Ah well, we only have 1 value per cell in this example. It might be easier to imagine the data-file as multivariate. For each combination of **Person** and **Condition** we only have 1 value. In order to test the interaction we need some variation, and our variation of one variable on each level of the other is 0. Look at the data below, when we look at [PP=1; CONDITION=1]=[M=37; SD=0]. That's it, just one value with zero variation.

PP	Value	Condition	PP	Cond_1	Cond_2	Cond_3
1	37	1	1	37	29	58
1	29	2	2	54	66	89
1	58	3	3	71	89	51
2	54	1				
2	66	2				
2	89	3				
3	71	1				

With zero variation the math kinda breaks down, dividing by zero at some points, interaction cannot be computed with just one value per cell. This is why the **Condition** and **Person** factors have the same **Error**, the error of the interaction cannot be separated from the overall error.

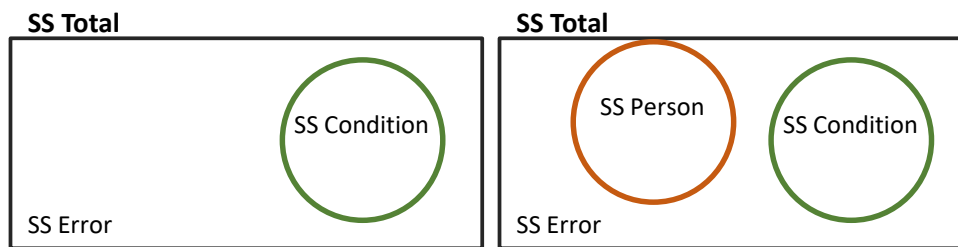
Box 2. What does it mean that Interaction cannot be separated from error?

Think of the interaction of **Time-point** and **Person** like this. We want to know if there is a difference between T1, T2, and T3 for Person 1 (a **Time** effect per level of the **Person** effect). We'd compare three individual values with 0 variation/variance (since each **Person** only has 1 value per **Time-point**). It's impossible to know if the differences are because of the **Time** effect or because of random events. We can't determine what is caused by **Person** specifically and what is unexplained by various other factors.

If there are multiple factors then the situation changes. If we would add another factor to this design where each participant did the test several times in each condition (**Replication**) then there would be multiple variables in each cell. For the **Time** effect we would have say 5 values (one for each **Replication**), giving us variation to work with. For the **Replication** effect we would have 3 values (one for each **Time-point**). For the Three-way interaction of **Time*Replication*Person** we'd get stuck again and can't separate it from the **Error**.

1.2.3 Effect of adding Person as random Variable

Adding **Person** as a random factor effectively informs the model that we are dealing with a WS design, it tells the model that those 120 people are actually the same 40 people three times. This means that a part of the variation in the data can be explained by the fact that it's the same person. That two data-points between Time-point 1 and 2 are both very high isn't just random chance, it's because it's the same person who happens to be very good. We are now explaining more (smaller error) and our degrees of freedom become lower. We are dividing a much smaller SSE by a smaller DF leading to a larger F.



$$MSE = \frac{SSE}{DFe} = \frac{42,223}{117} = .361 \quad F = \frac{MSG}{MSE} = \frac{2,037}{0,361} = 5.643$$

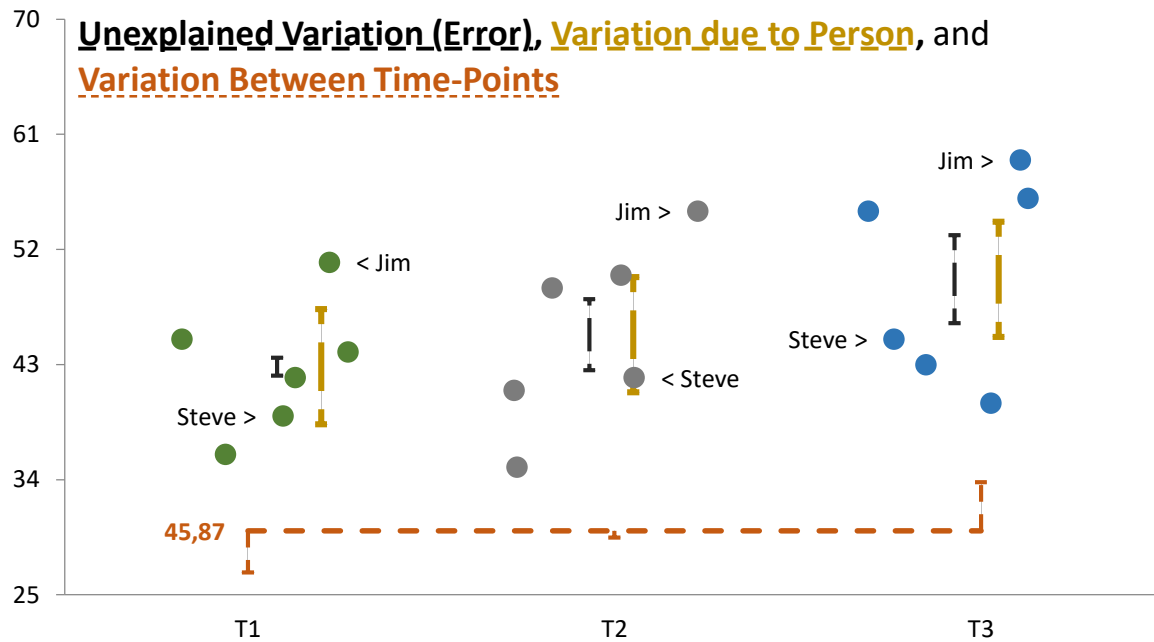
$$MSE = \frac{SSE}{DFe} = \frac{6,362}{78} = .082 \quad F = \frac{MSG}{MSE} = \frac{2,037}{0,082} = 24.969$$

Pooled Standard Deviation and Standard Error

The tests we used, the Univariate ANOVA with random PP, does come with a small caveat. Remember that in a normal ANOVA the variances must be equal in each group? This means that MSE is an **Unbiased Estimator of the error variance** (we expect it to only express error). If the assumption is violated the estimation will be off because it will be biased towards one of the groups (and then the whole logic of MSB/MSE doesn't work because MSE also has a group effect in it). Since we assume homogeneity we calculate the error in each group and then pool that together. That's why we have a *Pooled Variance*, a *Pooled Standard Deviation*, and a *Pooled Standard Error* (and why violation is a problem for your results).

When we add a random effect of **Person** there's still this implicit assumption that might not be obvious just yet. It assumes that the effect of **Person** is the same at each **Time-point**. It assumes that the variation caused by **Person** doesn't change, since we have the same people in each time-point/condition. So all the variation caused by person 1 being... well just themselves, will be the same regardless of **Time-point**. Everything that is left is unexplained is **Error**. This means that the MSE (of the PP*Condition effect) is again an **Unbiased Estimate of the error variance**.

It's more complicated than simply homogeneity of variances though. We assume that the effect of **Person** is the same, regardless of **Time-point**. This means that, yes, the variances of **Person** will be the same at each **Time-point**. But the real assumption is about the **Person** effect, you being you will have the same effect in each **Time-point**.



Looking at this image again it might become clearer how this implicit assumption of equality comes about, and why we cannot simply assume this. We pick one person (Jim), indicated by the arrow and the obvious label. Jim has a high score on whatever scale we have. Since we have the same people in each **Time-Point** Jim has a consistent high score, the highest in all **Time-Points**. The model assumes that the **Person** effect is the same in each **Time-Point**. Jim will be 10 points higher than one other specific person, we call him Steve. Because we assume the **Person** effect is the same in each **Time-Point** Jim will always score 10 points higher than Steve.

In some cases this might make sense, but not always. Imagine that this is a drug treatment study. We start at T1 where nobody got any medication, and in T2 they did get medication. Currently the model assumes that the effect of the medication will be the same for each **Person** (no interaction). In reality though there is so much going on that it's reasonable to imagine that in some people the drug will do more or less than in other people, in other words we can expect interaction.

The current model, with **Person** as a random factor, assumes that the variation due to **Person** is the same in each **Time-Point**. With this implicit assumption the ANOVA and the Random Effect ANOVA use a **pooled error variance**. This can be seen in the Pairwise Comparisons. The Standard Error is based on the Pooled Variance, which in the ANOVA is MSE. The full formula for calculating t is:

$$t = \frac{\text{Difference}}{\text{Standard Error}} = \frac{\hat{Y}_1 - \hat{Y}_2}{\sqrt{S_{Pooled}^2 \left(\frac{1}{N_1} + \frac{1}{N_2} \right)}} = \frac{\hat{Y}_1 - \hat{Y}_2}{\sqrt{MSE \left(\frac{1}{N_1} + \frac{1}{N_2} \right)}}$$

t is the difference divided by the Standard Error, meaning the lower part is the formula for the Standard Error: $\sqrt{MSE \left(\frac{1}{N_1} + \frac{1}{N_2} \right)}$. For the first ANOVA this number comes down to $\sqrt{0.361 \left(\frac{1}{40} + \frac{1}{40} \right)} = 0.134$ and this standard error is used in all of the pairwise comparisons.

Pairwise Comparisons

Dependent Variable: Value

(I) CONDITION	(J) CONDITION	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
1	2	.447*	.134	.001	.181	.713
	3	.169	.134	.212	-.097	.435
2	1	-.447*	.134	.001	-.713	-.181
	3	-.278*	.134	.041	-.544	-.012
3	1	-.169	.134	.212	-.435	.097
	2	.278*	.134	.041	.012	.544

Based on estimated marginal means

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

But we've overestimated the error in the first analysis, SSE and DFe was too big which led to an MSE that was too big. We fixed that in the second analysis by adding the **Person** effect. The Standard Error in this second analysis is $\sqrt{0.082(\frac{1}{40} + \frac{1}{40})} = 0.064$, which is a lot less than the 0.134 we had before. The result is that our confidence intervals shrink and the differences suddenly become more significant, even after adjustment for multiple comparisons. But we still assume that MSE is an unbiased estimator of the error variance, because we assume that the **Person** effect is equal in each **Time-Point**. The analysis runs with it and uses a pooled variance for the T-test.

Pairwise Comparisons

Dependent Variable: Value

(I) CONDITION	(J) CONDITION	Mean Difference (I-J)	Std. Error	Sig. ^a	95% Confidence Interval for Difference ^a	
					Lower Bound	Upper Bound
1	2	.447*	.064	.000	.320	.574
	3	.169*	.064	.010	.042	.296
2	1	-.447*	.064	.000	-.574	-.320
	3	-.278*	.064	.000	-.405	-.151
3	1	-.169*	.064	.010	-.296	-.042
	2	.278*	.064	.000	.151	.405

Based on estimated marginal means

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

1.2.4 Sphericity Assumption

We've added **Person** as a random factor, we've fixed the overestimation of SSE and DFe, the pairwise comparisons look good, so that's that? Yes and no. Remember that implicit assumption that led to the use of the **Pooled Standard Error**? Where we assume that the **Person** effect is the same for each **Time-Point**? Yeah... we didn't check for that. The Second Analysis is valid, but only in the case that this assumption is met (which it rarely is).

The assumption we're talking about is the assumption of **Sphericity**. This basically says that the variances of the difference scores are equal; if they are we can use the pooled Standard Error. In the second Analysis we didn't test for this because we're technically still doing a Between Subject ANOVA as far as the model is concerned (Condition and **Person** are both seen as BS effects, but since we correct for **Person** in the Condition effect it's all good).

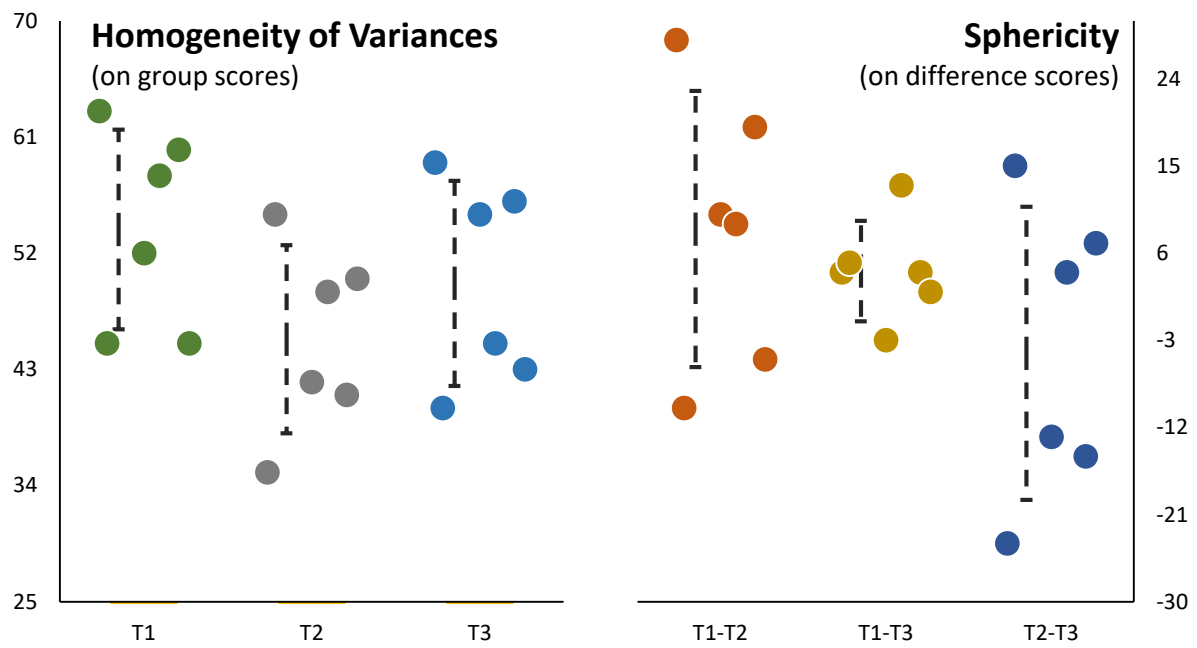
Sphericity is not the same as *homogeneity of Variances*, but it's similar. Let's look at an exaggerated example below.

PP	T1	T2	T3	T1-T2	T1-T3	T2-T3
1	63	35	59	28	4	-24
2	60	41	56	19	4	-15
3	45	55	40	-10	5	15
4	52	42	55	10	-3	-13
5	45	50	43	-5	2	7
6	58	49	45	9	13	4
Variance (Sample)						
	59.77	53.07	63.07	203.50	26.97	229.47
Variance						
(Population)	49.81	44.22	52.56	169.58	22.47	191.22

In the table we can see the difference between **Homogeneity of Variances** and **Sphericity**. In **Homogeneity of variances** the variances in each group needs to be equal (in case that wasn't obvious). That means that 59.77, 53.07 and 63.07 can't differ too much. **Sphericity** doesn't care about the groups itself; it cares about the differences between the groups. With three groups we can make three pairs. All of these pairs have difference scores for each person, and the variances on THOSE scores needs to be equal (203.50 vs. 26.97 vs. 229.47).

Whatever we did between T1 and T2 had a curious effect. The person with the highest score at T1 had the lowest score in T2. All scores in fact have reversed. The absolute differences are still the same: Steve and Jim still differ by 10 points so the variances will be the same. What the model assumes is that Steve will always score lower (technically it assumes that the change from T1 to T2 is the same for everyone, but details), and that's violated. We check this using the difference scores.

In the graphs below it might look better. Here we see the raw scores on the left with a line indicating the standard deviation (Variance is just Std.Dev² but those lines would be too large). The left graph shows homogeneity of Variances. On the right the difference scores are plotted along with the Standard Deviation of those difference scores. This shows **sphericity** and it should be obvious that in this example there is *homogeneity of variances*, but no **sphericity**.



1.3 Repeated Measures ANOVA

Since we can't simply assume **Sphericity** we would like to have a test for this. So instead of doing a Univariate ANOVA and cheat a little bit by adding **Person** as a random factor, we'll do a proper **Repeated Measures ANOVA**. We first need to restructure our data from *Univariate* to *Multivariate*, now each person only has one row.

PP	Value	Condition	PP	Cond_1	Cond_2	Cond_3
1	37	1	1	37	29	58
1	29	2	2	54	66	89
1	58	3	3	71	89	51
2	54	1				
2	66	2				
2	89	3				
3	71	1				

SPSS will give you more output than you need. Long ago the powers that be at IBM decided that they would give you all the tables, it's up to you to decide what you want to use. When you use a multivariate design, SPSS will know you mean that the **Person** effect is an actual person. In essence it will then add this **Person** effect itself, knowing it'll need to check **Sphericity**. That's nothing mathematical or statistical, more of a programming and design issue. Multivariate files are simply more intuitive for humans to use and tell the program that the number of people is equal to the number of rows.

That's why the RM ANOVA is sometimes called a multivariate test; multivariate refers to the data-file and not the type of test.

1.3.1 Multivariate Test

The first table we see is the Multivariate Tests Table. This is a test on its own and treats the data... well as a multivariate set (it's literally a MANOVA). In words this means that it sees each column (each Condition/Time-Point) as a separate independent test, while acknowledging that comes from the same person. In other words, the multivariate test includes the dependence/correlation due to the **Person** effect, but it ignores the dependency/correlation caused by the fact that it's the same person doing the same test. The Null hypothesis of this test is: $\mu_1 = \mu_2 = \mu_3$. More advanced, the multivariate test has no restrictions on the covariance matrix and estimates all the variances and covariances separately (Unstructured). We'll get back to covariance matrices later.

Multivariate Tests ^a						
Effect		Value	F	Hypothesis df	Error df	Sig.
COND	Pillai's Trace	,599	28,417 ^a	2,000	38,000	,000
	Wilks' Lambda	,401	28,417 ^a	2,000	38,000	,000
	Hotelling's Trace	1,496	28,417 ^a	2,000	38,000	,000
	Roy's Largest	1,496	28,417 ^a	2,000	38,000	,000
	Root					

a. Exact statistic

Because it ignores that it comes from the same test, treats it as independent measurements, and only takes the **Person** effect into account, there is no assumption of **Sphericity**. There is however an assumption of normality of differences scores that needs to be met. The multivariate test comes with four estimates of the effect; we usually look at *Pillai's Trace* or *Wilks' Lambda*.

1.3.2 Univariate Tests

The second part of the Repeated Measures ANOVA gives us the Univariate Tests. This is the one we prefer since it has more power than the multivariate test. The Univariate test recognizes both the **Person** dependency and the measurement dependency. It takes into account that we have the same people doing the same test multiple times.

But with this dependency included we get **sphericity** in the package. The variance of the Difference Scores needs to be equal in order to proceed. Mauchly's Test will provide us with the answer which in this case is...yes there is **sphericity**. The Null Hypothesis is that there is **sphericity** and we do not reject it. Unfortunately we can't really trust this test as it is very sensitive to violations of normality. It's better to always assume **Sphericity** has been violated.

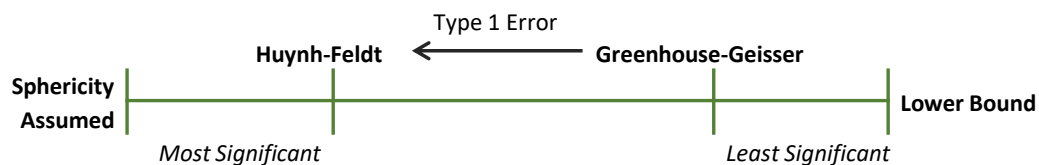
Mauchly's Test of Sphericity^a

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.
COND	.891	4.364	2	.113

Mauchly's Test of Sphericity^a

Within Subjects Effect	Epsilon ^b		
	Greenhouse-Geisser	Huynh-Feldt	Lower-bound
COND	.902	.943	.500

A violation of Sphericity isn't the end though. We know we have made an error on the estimate, so we can adjust for the fact that we might be making a Type-1 error. That's where the Epsilon Adjustment is for. The epsilon adjustment changes the degrees of freedom, changing the critical F-value and thus changing the significance. The epsilon adjustment is also based on the Covariance matrix, whole lot of math stuff.



The epsilon correction is applied to the degrees of freedom in the Tests of Within Subject Effect. Looking at Greenhouse-Geisser we can do the calculations by hand. That's all there is too it, we've successfully corrected for the inflated Type-1 error.

$$DF_{Cond|GG} = 2 * 0.902 = 1.804 \quad DF_{Error|GG} = 78 * 0.902 = 70.365$$

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
COND	Sphericity Assumed	4.073	2	2.037	24.969	.000
	Greenhouse-Geisser	4.073	1.804	2.258	24.969	.000
	Huynh-Feldt	4.073	1.886	2.159	24.969	.000
	Lower-bound	4.073	1.000	4.073	24.969	.000
Error(COND)	Sphericity Assumed	6.362	78	.082		
	Greenhouse-Geisser	6.362	70.365	.090		
	Huynh-Feldt	6.362	73.573	.086		
	Lower-bound	6.362	39.000	.163		

The Sphericity Assumed row is exactly the same result as you would get if you would have done a univariate ANOVA with **Person** as a random factor. They both assume sphericity, but the RM ANOVA tests this assumption and corrects for it.

1.3.3 Pairwise Comparisons

A difference between the previous examples and the repeated measures is the pairwise comparison. It might be obvious that they don't use a pooled Standard Error. For the comparison between groups the raw standard error is computed and used. Why? Because SPSS can't know which adjustment (or if any) you need or which test you want. So you get these and it's up to you to decide.

If we would use the pooled Standard Error we have to assume sphericity and that gives us:

$\sqrt{0.082(\frac{1}{40} + \frac{1}{40})} = 0.064$. We get the MSE from the **Tests of Within Subjects table** and will you look at that, it's the same as a Univariate ANOVA with **Person** as a random factor.


This also shows that **Pooled Standard Errors** will either over- or under-estimate the actual standard error. It's an average, so the true Standard Errors will hover around this mean. We see that some Standard Errors are bigger than 0.064 and other smaller. These pairwise comparisons are thus equal to Paired Samples T-tests.

Pairwise Comparisons

Measure: MEASURE_1

(I) COND	(J) COND	Mean Difference (I-J)	Std. Error	Sig. ^b	95% Confidence Interval for Difference ^b	
					Lower Bound	Upper Bound
1	2	.447*	.060	.000	.326	.567
	3	.169*	.057	.005	.053	.284
2	1	-.447*	.060	.000	-.567	-.326
	3	-.278*	.074	.001	-.427	-.129
3	1	-.169*	.057	.005	-.284	-.053
	2	.278*	.074	.001	.129	.427

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

An aerial photograph of a city at sunset. The sky is filled with dramatic, orange and yellow clouds. The city below is mostly in shadow, with some buildings and a large industrial area visible. A white rounded rectangle with a thin black border is positioned in the lower-middle part of the image, containing the title text.

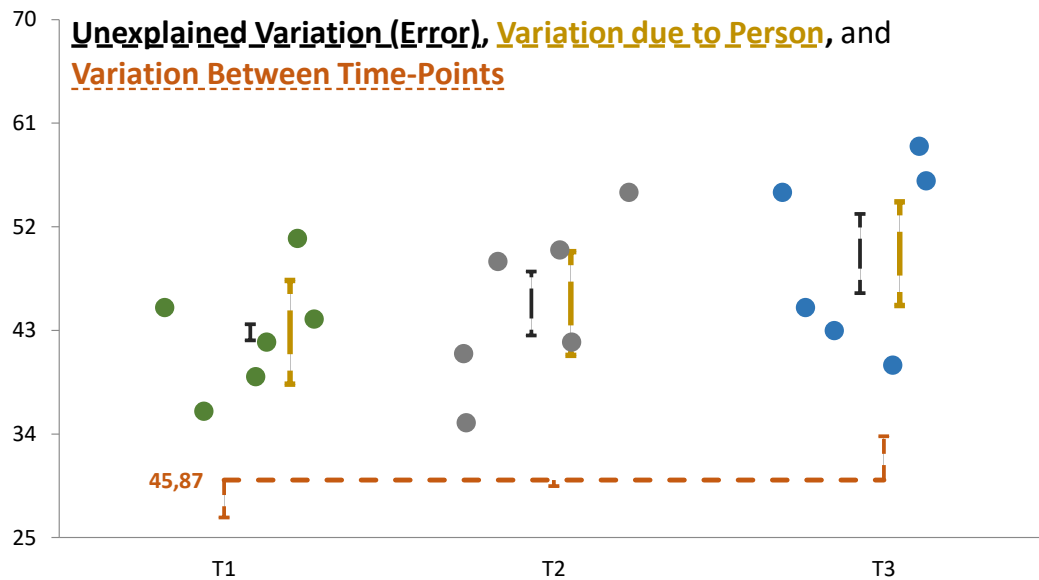
2. Mixed Models Procedure

Introduction

The second section of Session 1 covers the Mixed Models Procedure (in SPSS). We'll look at why you'd want to use Mixed Models over Repeated Measures ANOVA and then replicate a RM ANOVA using Mixed Models. The SPSS output of the Mixed Procedure will be discussed using an example.

Short Recap (in case you skipped the last chapter)

When discussing the Repeated Measures ANOVA we saw that it's actually just an ANOVA with a random **Person** effect shoved in there. Because of the **Person** effect we assume quite a bit, which is why we need to do some extra tests for **Sphericity** and adjust in case of the (likely) violation of that assumption.



The average of Time-point 1 is lower than the overall average because:

- 1) People in at Time-Point 1 scored lower than average (the time effect of interest).
- 2) Error due to factors not controlled by the experiment.
- 3) These specific people are in Time-Point 1 (some people are simply better/worse on this task).

Why does this specific person have a lower/higher score than the average within Time-point 1?

- 1) Error due to factors out of our control.
- 2) The person effect.

2.1 RM ANOVA versus Mixed Models

To determine the effect of time-point you could run a RM ANOVA, and then interpret the results from the Greenhouse-Geisser corrected Test of Within Subject effects. But let's be honest here for a bit, the RM ANOVA is becoming a fossil, it's antiquated. When comparing RM ANOVA and Mixed Models the results can sometimes be exactly the same, but in some cases they might differ vastly. Which to use?

RM ANOVA has one thing going for it: simplicity. Mixed Models are complicated, but so much more powerful. RM ANOVA is never better or more accurate than Mixed Models; at best they are just as good. The reason is simple, Mixed Models is a more general procedure that allows you to tweak more parameters and answer more sophisticated questions. RM ANOVA is bound by assumptions that you need to check and deal with, Mixed Models less so.

2.1.1 The Perfect Design

If you have a simple experiment like a pre-post measurement, it might even include a between subject factor, no missing data, and normal residuals... then you can safely use a RM ANOVA. The Mixed models will be more or less identical anyway and all assumptions are met. As nice as that sounds, we rarely find such perfect designs and perfect data outside of statistics courses.

2.1.2 Clustering or Multiple Levels

We deal with a lot of repeated measures and usually these are over time or space. You can do the same test several times, stimulate different areas, or have multiple conditions. These types of repeated measures can be handled by a RM ANOVA as well. The trouble comes when we introduce multiple levels: students within classrooms, subjects within countries, cells within dishes, and so on.

RM ANOVA can't take clustering into account. If you are lucky in your design you can turn it into a factor, but this isn't quite the same. RM ANOVA won't take into account the correlation between individuals within a group, but Mixed Models will. If your repeated measures are on the same level, RM ANOVA is going to be impossible.

The example below shows what I mean. The univariate data is on the left, where we have people in different groups (think classes, countries, or PBL groups) doing some kind of test/experiment three times. We don't care about the groups, but we do know which participants are together in which groups.

Because we don't want any confounding we balance out the order of the conditions. This is an issue for RM ANOVA, because order and condition are on the same level. Person is within groups, condition is within person, but order is also within person. Because Order and Condition are on the same level it's impossible to make a multivariate file that can check the Condition effect while also taking into account the order effect.

One might think that we can ignore group and go with a Condition-Order file. It sounds like a plan, but your RM ANOVA will not run because not all participants have data for all repeated measure. Participant 1 will only have data for [Condition=1; Order=1], while participant 2 only has data for [Condition=1; Order=2]. The same will happen for the other conditions, leading to all participants

being removed due to incomplete data. To make this run you will have to average over order, to check the condition effect and average over condition to check the order effect. It's possible and I'm sure this has been done in the past, but why would you if you don't have to? Our star Mixed Models doesn't care about any of that. It works with univariate files and will give you Condition, Order, and Condition*Order (if you have enough participants) effects, while also taking into account that people within groups have higher correlations than people between groups.

Group	PP	Condition	Order
1	1	1	1
1	1	2	2
1	1	3	3
1	2	1	2
1	2	2	3
1	2	3	1
2	3	1	3
2	3	2	1
2	3	3	2
2	4	1	1
2	4	2	2
2	4	3	3
3	5	1	2

PP	Cond1	Cond2	Cond3
1	45	--	--
2	--	32	--

2.1.3 The Case of the Missing Data

No matter how well we design our experiments or collect our data, data inevitably goes missing. A participant might not be available for the last session, their data might not be useable, or they might be an outlier on a single point.

I've already touched upon this in the previous paragraph. RM ANOVA will throw out your entire participant if they have a single missing value. You could test them 100 times, mess up once and all that data is gone. It's strict but necessary for the RM ANOVA math and procedures to work.

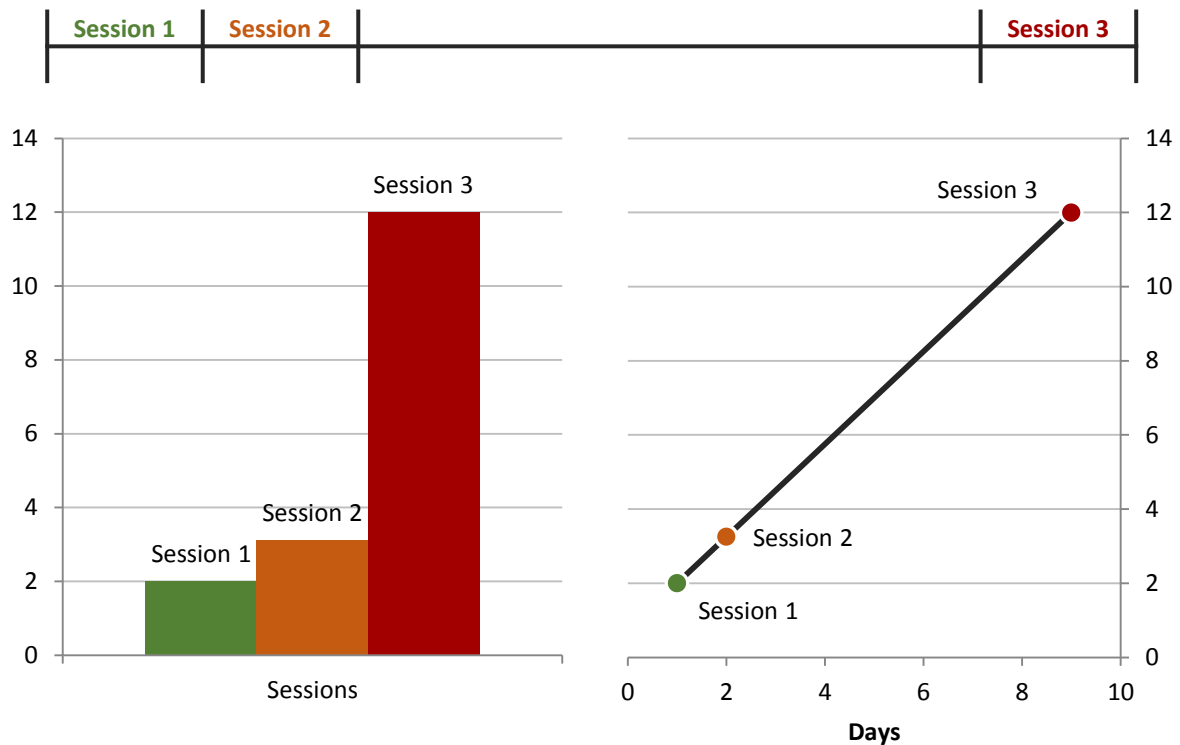
Mixed models on the other hand, ah they don't care about some data missing here and there. Participants aren't removed, single observations are. That participant who didn't complete the last session still has useable data for all others. Mixed Models will simply adjust and use whatever data is available. You do have to be careful that data isn't missing in some kind of pattern. If you have two groups, say treatment and control, and half of the people from the treatment condition didn't make it to the end you have a problem with your data (and with your treatment). Missing values should be missing at random to not bias the results.

2.1.4 The Factor of Time

Time is always a factor; in repeated measures this is pretty literal. A RM ANOVA will always treat your repeated measures as categorical factors. In a lot of designs this works, but in quite a few with more than 2 replications it might not. Imagine testing in multiple sessions: Session 1 on day 1, Session 2 on day 2, and session 3 a week later.

RM ANOVA will see these sessions as categorical, with the same distance between them. That might not be the best way to go about this design though. Mixed Models would allow you to put session

in as a continuous variable, where you can create a regression line. On top of that, mixed models would also allow you to specify this pattern in the repeated measures by using a certain covariance matrix.



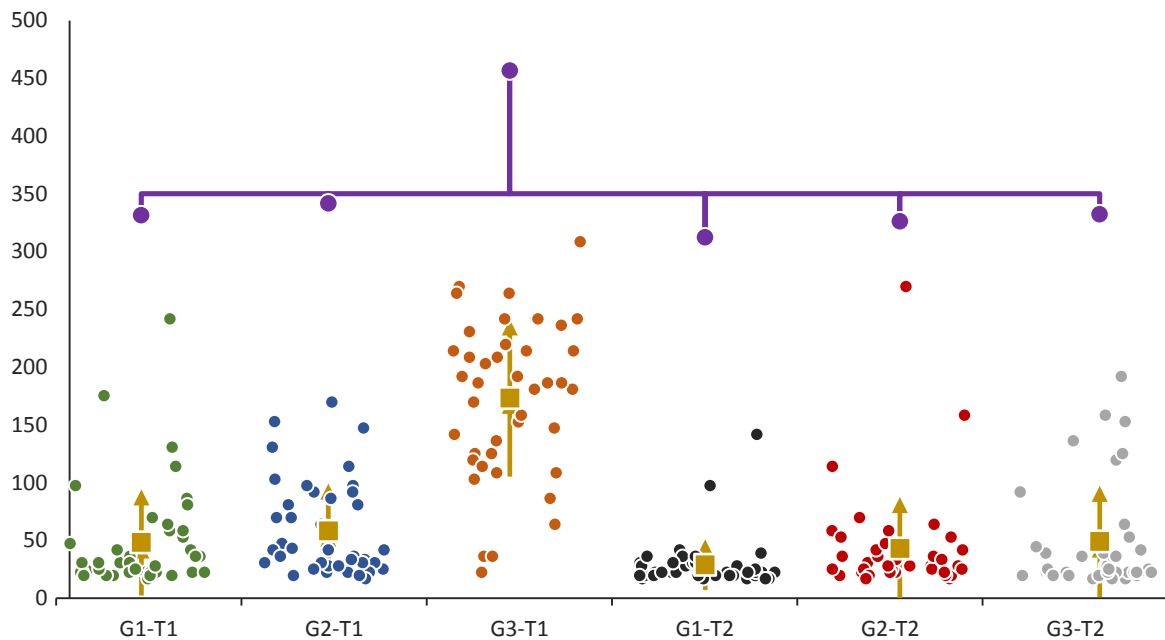
2.1.5 Winner: Mixed Models

These are some relevant reasons for our kind of work that show when Mixed Models are better than RM ANOVA. There are more reasons, some more complicated than others, but they all show the same thing: Mixed Models wins.

2.2 Example Data

For the example I chose a slightly more complicated design, a crossed 2x3 within subject design. There are two time-points and we measure our participants in three different conditions. This more closely resembles the type of designs we use in our experiments.

The design is crossed instead of nested, which means that all participants go through all conditions and all time-points. Nested means that certain participants belong to one group (such as class) but not the others. I could've made this a nested design by testing some people during day 1 and some people during day 2 (then participants is nested in day), the people during day 1 might correlated more with each other than with the people at day 2.



Just looking at the data we can already expect some things. It seems that T1 has higher scores than T2, Group three also scores higher than the other two, and an interaction probably isn't out of the question either with the large group effect during T1.

2.3 Using Repeated Measures ANOVA: Output

First up is good old Repeated Measures ANOVA. We'll skip some of the steps and go to the relevant tables right away (most of this was covered in Chapter 1 anyway and I'm assuming some knowledge here).

Mauchly's test is pretty clear, we do not have sphericity for group or the interaction. Time isn't tested because it has two levels, so Sphericity is automatically assumed.

Mauchly's Test of Sphericity^a

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.
Time	1,000	,000	0	.
Group	,691	14,766	2	,001
Time * Group	,662	16,477	2	,000

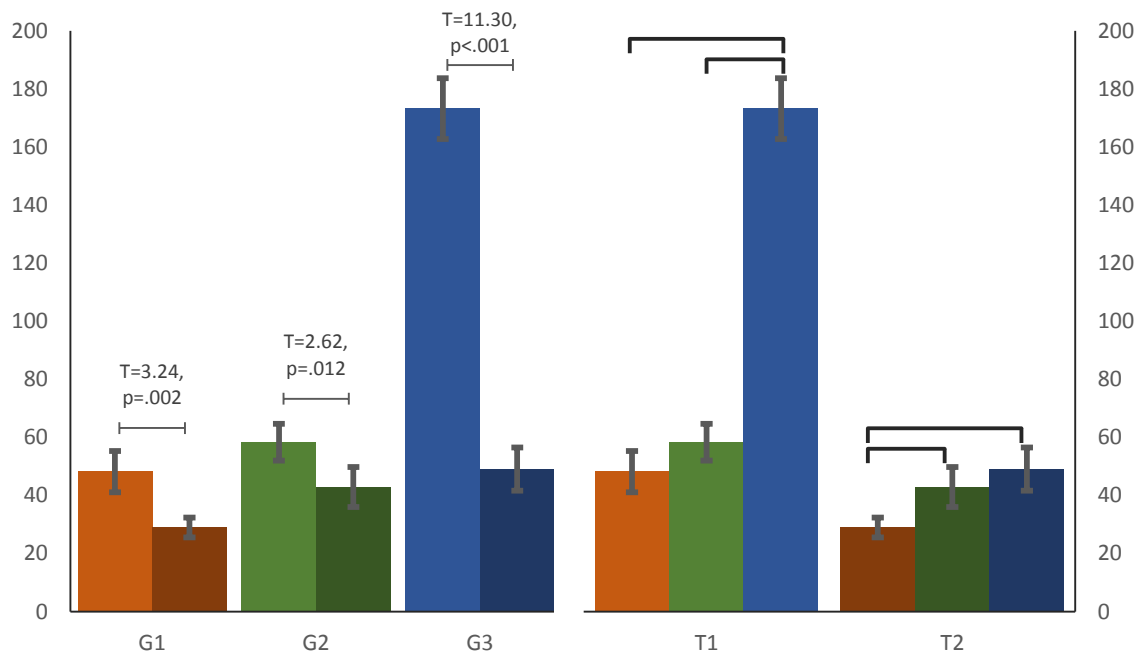
Epsilon		
Greenhouse-Geisser	Huynh-Feldt	Lower-bound
1,000	1,000	1,000
,764	,788	,500
,748	,769	,500

Looking at the Test of Within-Subject Effects we can see main effects for Time ($F_{(1,41)}=151.47, p<.001$), Group ($F_{(1.53,62.66)}=82.02, p<.001$), and their interaction ($F_{(1.50,61.30)}=56.58, p<.001$).

Tests of Within-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Time	176352,868	1,000	176352,868	151,475	,000
Error(Time)	47733,841	41,000	1164,240		
Group	254081,038	1,528	166254,728	82,023	,000
Error(Group)	127005,312	62,659	2026,935		
Time * Group	159956,084	1,495	106981,366	56,584	,000
Error(Time*Group)	115902,626	61,302	1890,675		

The interaction can be split up into a Time effect per group or a Group effect per time-point. You could make any contrast actually, comparing the groups to a control instead of each other for instance, but let's stick to the familiar unplanned comparisons. Those inform us we have Time effects in each group, which is nice. The other way around we see a group effect in both time-points. In T2 it's group 3 that is significantly higher than the other two, while in T2 it's group 1 that is significantly lower than the other two.



Pairwise Comparisons

Group	Mean Difference (T1-T2)	Std. Error	Sig. Sidak
T1	19,218*	5,924	,002
T2	15,373*	5,860	,012
T3	124,133*	10,983	,000

Pairwise Comparisons

Time	(I) Group	(J) Group	Mean Difference	Std. Error	Sig. Sidak
T1	G1	G2	-10,110	6,915	,389
		G3	-125,021*	12,483	,000
	G2	G3	-114,912*	12,004	,000
T2	G1	G2	-13,955*	3,852	,002
		G3	-20,106*	5,556	,002
	G2	G3	-6,151	5,454	,605

2.4 Using the Mixed Procedure: Output

Let's re-do the analysis, but use Mixed Models instead. The design isn't crazy complicated, no nested factors, so the results should be the same. To get the same results we'll use Mixed Models with an Unstructured Covariance Matrix for repeated measures. Unstructured is what RM ANOVA uses for the pairwise comparisons. More about covariance matrixes will follow, first we need to establish a good basis for using the procedure and reading the output.

The very first table SPSS gives us is the Model Dimensions table. If you're going to run multiple models it's smart to copy this to your Statistical Analysis Report. It tells us exactly what kind of model was run: A model with Fixed Effects Time, Group, and their interaction. Repeated Measures where Time*Group and the covariance matrix was Unstructured.

Model Dimension

		Number of Levels	Covariance Structure	Number of Parameters	Subject Variables	Number of Subjects
Fixed Effects	Intercept	1	Unstructured	1	PP	42
	Time	2		1		
	Group	3		2		
	Time * Group	6		2		
Repeated Effects	Time * Group	6	Unstructured	21	PP	42
Total		18		27		

This gives us a model with a total of 27 parameters to estimate: all the regression coefficients (6 values) and all the variances and covariances in the model (21 values).

For regression we do not want the number of parameters to come close to the number of observations, they need to be as far apart as possible. In this example it's not that much of a problem because we have 42 participants with six observations each ($42 \times 6 = 252$ observations). It is certainly possible to have more parameters than observations though, quite easily if you're using an unstructured covariance matrix.

The variances and covariances the model has to estimate now is $((6 \times 6 - 6)/2) + 6 = 21$. Adding one more group gives us $2 \times 4 = 8$ levels for the interaction, resulting in a structure with $((8 \times 8 - 8)/2) + 8 = 36$ parameters. Adding another time-point is even better, making it a $3 \times 3 = 9$ design and a 45-parameter structure matrix. Effects and levels add up fast, but more on that in later sessions when we cover covariance matrixes and model construction. The information criteria table tells us how well the model fits the data. On its own the numbers are meaningless and cannot be interpreted. They are going to be very useful later on when we start comparing model.

Information Criteria^a

-2 Restricted Log Likelihood	2402,637
Akaike's Information Criterion (AIC)	2444,637
Hurvich and Tsai's Criterion (AICC)	2448,762
Bozdogan's Criterion (CAIC)	2539,249
Schwarz's Bayesian Criterion (BIC)	2518,249

Then we get to the good stuff, the **Tests of Fixed Effects**. This is more like the overall test for a regression than the ANOVA table we're used to. It tells us we have a significant Time effect ($F_{(1,41)} = 151.47$, $p < .001$), Group effect ($F_{(2,41)} = 57.79$, $p < .001$), and interaction ($F_{(2,41)} = 38.68$, $p < .001$).

We don't need to adjust for violations for sphericity because this test has no assumption of sphericity. The procedure doesn't make assumptions about the data to make the math work, but you have to. In the session about model selection we'll go deeper into this, but it comes down to you deciding what the structure of the data is. If there is compound symmetry then you can fit a CS structure, if there's a different pattern you fit that pattern. There is no assumption because the test requires you to make that call.

Type III Tests of Fixed Effects^a

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	41,000	183,959	,000
Time	1	41,000	151,474	,000
Group	2	41,000	57,788	,000
Time * Group	2	41,000	38,684	,000

The parameter coefficients can be found in the **Estimates of Fixed Effects**. Take for example the [Time=1] effect, it shows as significant ($p < .001$), which means that [Time=1] is significantly different from the reference, which is [Time=2]. This is nice, but this table should be used for interpretation and less for making conclusions. The effects you see is a +1 change in the variable, for a categorical this means changing groups, when keeping the rest constant. To make inferences it's better to use **Estimated Marginal Means**, which takes all effects into account. The Parameter Estimates are more useful for covariates (continuous variables).

Estimates of Fixed Effects							
Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	49,074048	7,437278	41,000	6,598	,000	34,054161	64,093935
[Time=1]	124,133571	10,983209	41	11,302	,000	101,952530	146,314613
[Time=2]	0	0
[Group=1]	-20,106190	5,555638	41	-3,619	,001	-31,326029	-8,886352
[Group=2]	-6,150714	5,454389	41	-1,128	,266	-17,166076	4,864648
[Group=3]	0	0
[Time=1] * [Group=1]	-104,915714	13,819458	41	-7,592	,000	-132,824677	-77,006752
[Time=1] * [Group=2]	-108,760000	12,385732	41	-8,781	,000	-133,773493	-83,746507
[Time=1] * [Group=3]	0	0
[Time=2] * [Group=1]	0	0
[Time=2] * [Group=2]	0	0
[Time=2] * [Group=3]	0	0


Because we have an interaction we'll have to look at the effect of one per level of the other. In a plot-twist the likes you've never seen, these pairwise comparisons are identical to the ones we had for RM ANOVA.

Pairwise Comparisons

Group	Mean Difference (T1-T2)	Std. Error	Sig. Sidak
G1	19,218*	5,924	,002
G2	15,373*	5,860	,012
G3	124,133*	10,983	,000

Pairwise Comparisons

Time	(I) Group	(J) Group	Mean Difference	Std. Error	Sig. Sidak
T1	G1	G2	-10,110	6,915	,389
		G3	-125,021*	12,483	,000
	G2	G3	-114,912*	12,004	,000
T2	G1	G2	-13,955*	3,852	,002
		G3	-20,106*	5,556	,002
	G2	G3	-6,151	5,454	,605



3. Building a Model

Introduction

The last sessions focused on the **logic of Repeated Measures**; the influence of taking **Person** into account. We saw that **Person** improves the power of your tests and shrinks the standard error. **Repeated Measures ANOVA** automatically adds this **Person** effects, but keeps one big downside: the assumption of **Sphericity**. We assume that the **Person** effect is the same for each time-point and that's a bit silly. RM ANOVA does its best to fix this mistake by adjusting your tests using **Greenhouse-Geisser**, but yeah not really what we want.

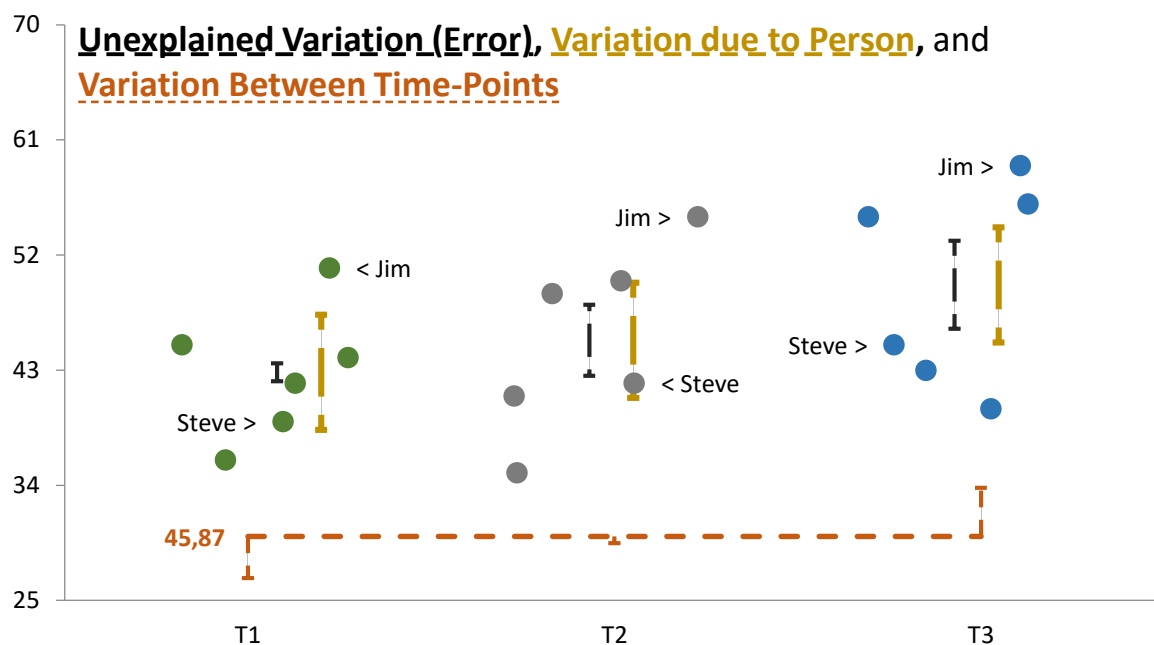
The **Mixed Procedure** was able to do the same things as the **RM ANOVA**, giving us the same results in the end. The best thing about **Mixed Models** is that they aren't limited to **Sphericity**, we can specify the pattern in the repeated measures. That's what we'll be doing in this session.

We will look at how to properly build a Mixed Model, go through some of the steps and choices you'll need to make. This session covers Variance-Covariance Matrices, Assumptions of the Model, Comparing Models, Model Reduction, and writing it all down.

3.1 Covariance Matrices

For repeated measures and mixed models we saw that it's a regression model with a person effect included. RM ANOVA and Mixed Models resulted in the same conclusion because they are pretty much the same model. The person effect we added in the RM ANOVA came with some assumptions though, the big one being Sphericity.

Sphericity meant that the variance of the difference scores are all equal, in other words the differences between people are the same in each time-point. This is a big assumption to make and in real life not very realistic. In RM ANOVA we acknowledge that it's ridiculous to assume this, so we try to adjust for it with the epsilon correction. But why? Why fix a mistake if you can prevent it? Why not do it right from the start? That's what Mixed Models allows you to do. One of the toughest things about Mixed Models is the covariance matrix, so we'll look at those first before we go into building a model.



The whole thing about repeated models is that we measured something from the same individuals multiple times. All of our measurements have a certain amount of variance, that's what we work with to do our stats. Because the measures are from the same test done by the same people we also have significant covariance, they are connected by an underlying dependency from a common source. One of the simplest ways of modeling that dependency, that correlation, is with **compound symmetry**.

Compound Symmetry means we assume that **the covariance between T1 and T2 is the same as the covariance between T1 and T3 which is the same as the covariance between T2 and T3**. It doesn't matter how far time-points are apart they will have the same covariance. The analysis will ignore any ordering of your data, all your pairs will be equal and exchangeable. In easier terms: **equal covariances means that people with a high score on T1 will have a high score on T2 and a high score on T3** (with positive covariances), the scores are dependent (they co-vary together).

Box 3. Tiny Mindblow

Covariance is a measurement of the linear relationship between two variables, which sounds a bit familiar doesn't it? Covariance is the raw score while correlation is the standardized form. This means that covariance can range from minus infinity to infinity, making it a bit difficult to say how strong the relationship is. Correlation on the other hand fits between -1 and 1, no matter what the scale of the variables is. Close to -1 or 1 means it's linear while closer to 0 means two variables aren't dependent on each other at all.

$$Covariance_{X,Y} = \frac{\sum_{i=1}^n (x_i - \hat{X})(y_i - \hat{Y})}{n - 1} \quad Correlation_{X,Y} = \frac{Covariance_{X,Y}}{SD_X * SD_Y}$$

The matrix form of compound symmetry is shown below. We have a covariance matrix, with the variances on the diagonal and the covariances on the offset. With compound symmetry all the variances are equal, as are all the covariances. Variance is then equal to σ^2 and the correlation is ρ . Only two values need to be determined which are σ^2 and ρ .

$$CovarianceMatrix : \begin{bmatrix} Var_1 & Cov_{1,2} & Cov_{1,3} \\ Cov_{2,1} & Var_2 & Cov_{2,3} \\ Cov_{3,1} & Cov_{3,2} & Var_3 \end{bmatrix} \quad CompoundSymmetry : \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

Sphericity is a different kind of structure but related. It checks the variances of the difference scores. But note that even *if compound symmetry isn't met we can still have sphericity*, **compound symmetry is sufficient, not necessary** (with CS there's always sphericity, without CS we should test for sphericity, see Box 4). There is one case where you always have **Sphericity** and don't need to test for it: if you have only two levels. If you only have two levels then you only have 1 difference score and **Sphericity** is always assumed. But here we have three levels, three time-points, so we can't just assume it.

How is **Sphericity** related to the covariance matrix? Other than using the difference score calculation you could use the formula below. If we use our previous example for T1 and T2 we get the same values as before when we calculated the the variance of the difference scores.

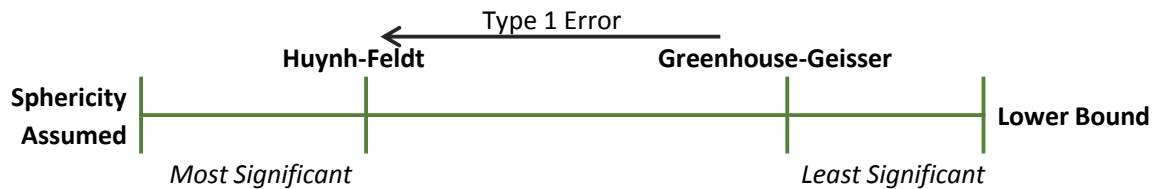
$$CovarianceMatrix : \begin{bmatrix} 59.77 & -45.33 & 47.93 \\ -45.33 & 53.07 & -56.67 \\ 47.93 & -56.67 & 63.07 \end{bmatrix} \quad \begin{aligned} S_{x-y}^2 &= S_x^2 + S_y^2 - 2(S_{xy}) \\ VarDiff &= Var_T1 + Var_T2 - 2(CoVarT1;T2) \\ &= 59.77 + 53.07 - (2 * -45.33) = 203.5 \end{aligned}$$

Box 4. Sphericity without Compound Symmetry?

The best way to show this is with an example. No matter which combination of Time-Points you choose, the variance of the difference scores will always be 20. In the left matrix (VarCovar1) we have perfect compound symmetry, the variances and covariances are all the same. In the right matrix (VarCovar2) all of the variances and covariances differ, so no compound symmetry but still perfect sphericity.

$$\text{VarCovar1} = \begin{array}{ccc|c} & T1 & T2 & T3 \\ \begin{bmatrix} 20 & 10 & 10 \\ 10 & 20 & 10 \\ 10 & 10 & 20 \end{bmatrix} & T1 \\ & T2 \\ & T3 \end{array} \quad \text{VarCovar2} = \begin{array}{ccc|c} & T1 & T2 & T3 \\ \begin{bmatrix} 10 & 5 & 10 \\ 5 & 20 & 15 \\ 10 & 15 & 30 \end{bmatrix} & T1 \\ & T2 \\ & T3 \end{array}$$

In the last session we talked about the consequences of adding a **Person** effect. In order for the math to work out and the logic of ANOVAs to hold, sphericity is automatically assumed. Rather than having all the covariances equal, Sphericity only requires you to have equality in the various pairs we can make with all conditions. How do we test for this? When you order a Repeated Measures ANOVA you get a table called Mauchly's Test of Sphericity. Mauchly gives you a value (W) which is calculated using the covariance matrix and is comparable to a Chi-Square. It has a distribution and by using the degrees of freedom it determines significance. The significance tests of the univariate test in the RM ANOVA always assume sphericity, when this assumption is violated the test made a mistake, showing us significance where there is none. The epsilon corrections are all calculated based on the covariance matrix as well, they represent the amount of violation, how much sphericity is violated. An epsilon of 1 means that the matrix has perfect sphericity, lower than 1 means it's off. Some even say that the epsilon correction is a better gauge for sphericity violation than Mauchly, so it's good to look at both.



When we use RM ANOVA we get a choice, either the general Unstructured matrix of the Multivariate Test or the Compound Symmetry (light) structure with Sphericity in the Univariate test. That doesn't give us much to work with and our data might be described better with a different pattern.

$$\text{Unstructured} : \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix} \quad \text{CompoundSymmetry} : \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

3.2 Covariance Matrices in Mixed Models

When working with repeated measures we assume that the replications have a covariance (or correlation if you prefer). We assume that scores at measurement 1 affect scores at measurement 2 and measurement 3. How exactly these covariances between repeated measures look, that's what the covariance matrix is about.

3.2.1 Compound Symmetry (Univariate method)

In Repeated Measures ANOVA we have a less strict version of Compound Symmetry, also known as the exchangeable correlation structure with constant variance. This means that each repeated measurement has the same variance and all pairs are equally correlated (and sphericity is assumed).

In the matrix on the right we see compound symmetry. The covariance between sessions is equal, no matter which pair you take (Session N vs Session N is of course 1). No matter how far apart the time-points are, the covariance/correlation is always the same, it's always Sigma (σ). In the Output below we see the output of Compound Symmetry. The diagonal offset is σ_1 and the Covariance is σ^2 . This structure assumes that all time-points have equal covariances and have interchangeable orders.

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho \\ & 1 & \rho \\ & & 1 \end{bmatrix}$$

Estimates of Covariance Parameters (using CS)

Parameter		Estimate	Std. Error	Wald Z	Sig.
Repeated Measures	CS diagonal offset	3,143316	,547181	5,745	,000
	CS covariance	-,166326	,231522	-,718	,473

3.2.2 First Order Autoregressive (AR1)

A very intuitive covariance structure, and often used, is the First Order Autoregressive structure. AR(1) assumes that the covariance is strongest for adjacent replications and systematically decreases with distance. The covariance between Rep 1 and Rep 2 is ρ_{1-2} , the covariance between Rep 1 and Rep 3 would be smaller ρ_{1-3} , and for Rep 1 and Rep 4 even smaller ρ_{1-4} .

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ & 1 & \rho & \rho^2 \\ & & 1 & \rho \\ & & & 1 \end{bmatrix}$$

In the matrix on the left we see the First Order Autoregressive structure. As the replications are further apart the covariance decreases systematically. From ρ to ρ^2 to ρ^3 etc. Do note that AR(1) is only applicable for **evenly spaced intervals** in the Repeated Measure. The output is similar to Compound Symmetry, giving us a diagonal and a rho value.

Estimates of Covariance Parameters (using AR1)

Parameter		Estimate	Std. Error	Wald Z	Sig.
Repeated Measures	AR1 diagonal	2,977826	,451468	6,596	,000
	AR1 rho	-,084592	,125749	-,673	,501

3.2.3 Toeplitz (TP)

A more complex structure is the Toeplitz Covariance Matrix. In this structure every diagonal is equal. The covariance between 1 and 2 is the same as the covariance between 2 and 3. Similarly the covariance between 1 and 3 is the same as the covariance between 3 and 4. Covariances are dependent on the distance between elements, but pairs with equal distances have the same covariance. The Covariance Structure works well with Ordinal data, where it makes sense that neighboring data-points have similar covariances.

$$\sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ & 1 & \rho_1 & \rho_2 \\ & & 1 & \rho_1 \\ & & & 1 \end{bmatrix}$$

Estimates of Covariance Parameters (using TP)

Parameter		Estimate	Std. Error	Wald Z	Sig.
Repeated Measures	TP diagonal	93,884458	21,233791	4,421	,000
	TP rho 1	0,811415	0,045958	17,656	,000
	TP rho 2	0,758609	0,059487	12,753	,000
	TP rho 3	0,728157	0,066433	10,961	,000

3.2.4 Unstructured (Multivariate Method)

$$\begin{bmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\ & & \sigma_3^2 & \sigma_{43} \\ & & & \sigma_4^2 \end{bmatrix}$$

Unstructured is the most complex structure, it puts no constraints on the matrix and gives the raw variances (diagonal) and covariances (offsets). These come back in the output, where (1,1) stands for Column 1 Row 1. This is the multivariate test in the RM ANOVA, which is pretty much a MANOVA that ignores any equality between the various measures.

Estimates of Covariance Parameters (using Unstructured)

Parameter		Estimate	Std. Error	Wald Z	Sig.
Repeated Measures	UN (1,1)	3,597120	1,084572	3,317	,001
	UN (2,1)	,008515	,688305	,012	,990
	UN (2,2)	2,897519	,873635	3,317	,001
	UN (3,1)	-,159329	,651767	-,244	,807
	UN (3,2)	-,082851	,584435	-,142	,887
	UN (3,3)	2,591025	,781223	3,317	,001
	UN (4,1)	,191760	,680538	,282	,778
	UN (4,2)	-,314204	,613351	-,512	,608
	UN (4,3)	-,641847	,592553	-1,083	,279
	UN (4,4)	2,822296	,850954	3,317	,001

3.3 Which Structure to use?

When deciding on your model it's important to know which model has the best fit. Unfortunately this isn't as easy as checking a single number, it requires some thinking and some insight into your model. Usually you will try different covariance structures based on what is theoretically possible in your data. For data that is recorded at different time-points an AR(1) structure is possible, while it makes less sense to use such a structure if you tested someone on the same day in different conditions. The unstructured matrix will always give you the best fit, because it imposes no restrictions on the covariance matrix. This makes it tempting to always use Unstructured, but it comes with some downsides. The first and most obvious downside is over-fitting. Your model fits your data very well, but only your data. When applied to a different set it might fail, making it a bit useless in the real world. That's not the only problem, the Unstructured matrix also requires a lot of degrees of freedom, probably more than you can afford unless you have a lot of data and a lot of time to model it. A small 2x2 design only has 3 parameters to estimate, two variances and their covariance. But these numbers increase dramatically (with a 4x4 matrix you're already up to 10).

Structure	Repeated Measures					
	2	3	4	5	6	7
Unstructured	3	6	10	15	21	28
Heterogeneous TP	3	5	7	9	11	13
Heterogeneous AR(1)	3	4	5	6	7	8
Heterogeneous CS	3	4	5	6	7	8
TP	3	3	4	5	6	7
AR(1)	2	2	2	2	2	2
CS	2	2	2	2	2	2

3.3.1 When Unstructured Doesn't work

The most common use for covariance matrices is in models that estimate repeated measures or longitudinal data. Here the variances and covariances represent each person's non-independent residuals. The data is connected, correlated, because it's the same person providing the data at different time points. More often than not you will see patterns in the variances and covariances; they might be nearly equal for instance. You will want to impose those constraints by setting up a compound symmetry structure, saving you a lot of degrees of freedom without having to give up a whole lot of fit. As mentioned before, if you have a lot of observations and only a few participants an Unstructured Matrix isn't the way to go. With three observations you will have to estimate 3 variances and 3 covariances, but with six this increases to 6 variances and 15 covariances. Remember that repeated measures include interactions, so a 2x3 design already has 6 repeated measures and 21 values to estimate. If you only have 20 people to estimate 21 values, you'll run out of degrees of freedom if you try an Unstructured Matrix.

3.3.2 When Unstructured Does work

Unstructured matrices are best when there's no pattern in your covariances. Most designs do come with patterns; humans can be quite consistent like that. This isn't the case for random effect though; random effects like random intercepts or random slopes usually don't show much consistency. Every person is different, which is why we add these random effect in the first place. Unlike the fixed effect, random effects often benefit from an Unstructured matrix. These random matrices (G-matrix) aren't all that big in most cases. Often only a single random intercept and slope needs to be estimated, which works fine for an Unstructured matrix. When dealing with random effects, in most cases, you'll choose between Unstructured or Variance components (which fits the variances but sets covariances to 0).

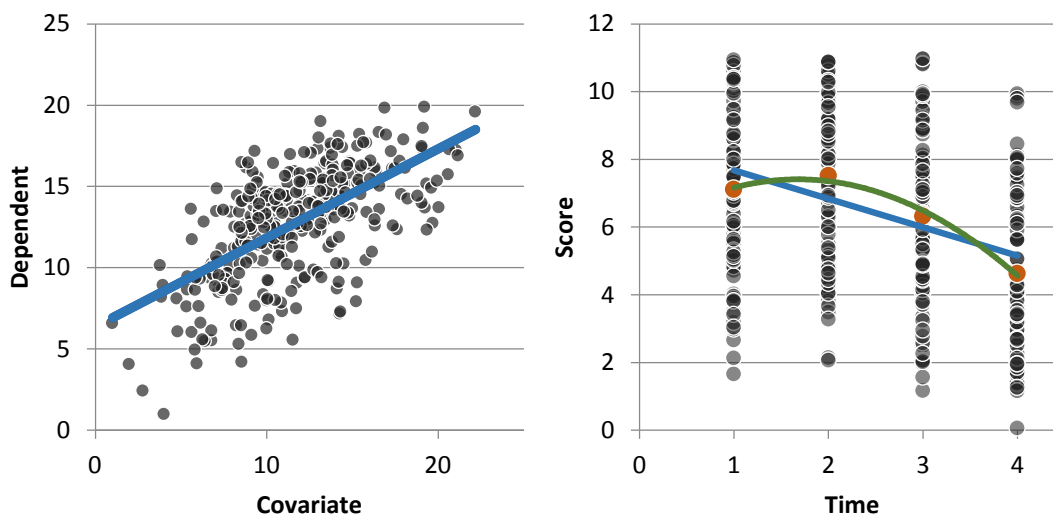
3.4 Assumptions of Mixed Models

The assumptions of Mixed Models are more or less the same as that of regular regression models. Depending on your data (is it multi-level? Are there covariates?) more assumptions might need checking, but for the majority of models a few are important.

3.4.1 Linearity

As fancy and complicated it all is, we are still using linear models and they assume linear relationships. If your model contains covariates (that is continuous predictors/independents) then those need to show a linear relationship with the dependent.

How to check this? Scatter-plots are your friend here. Alternatively you could also plot the residuals of the model against the observed scores to see if a pattern arises (if there is a pattern you missed a higher order effect).

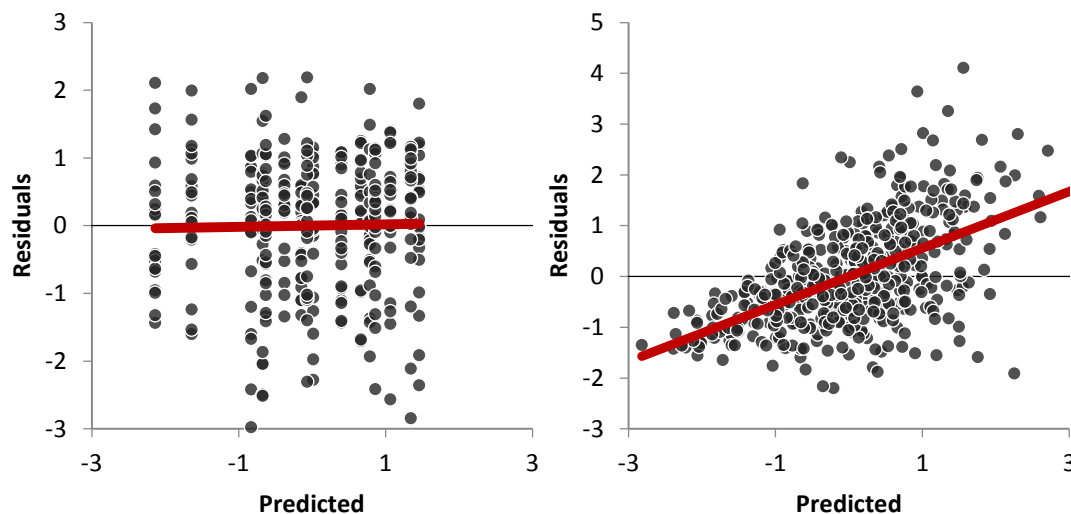


What if it's violated? The graph on the right looks linear, but we can see that a second order polynomial fits better. This isn't strange, especially with certain learning experiments. You could add time as a factor, but that's not the solution we were looking for. Instead you run a non-linear analysis. You can add a Time*Time factor to take into account the non-linear relationship between time and score. Alternatively you can transform the data, see if that fixes the issue.

3.4.2 Homoscedasticity

Homoscedasticity (or its opposite: Heteroscedasticity) has to do with your model and how well it can predict based on the information it has. Homoscedasticity means that the variance is the same for all your predicted values. Violation means that certain values have more or less variance. Because regression tries to minimize residuals, those with large variances will have "more pull" biasing your model. A biased model in turn leads you to conclusions that might not be generalizable.

How to check this? Scatter-plots again! This time we plot residuals and predicted values (use standardized predicted and residuals to also check outliers). The graph should show us a rather random distribution of points, meaning there is no pattern in the residuals. For those who don't like the visual interpretation, we can also do a more mathematical test. It's easier in R but works in SPSS too. We want to make sure nobody has larger residuals; an easy way to check this is to do an ANOVA on the Absolute residuals squared (using ID/PP as predictor). If that ANOVA shows no significant differences then you're good).



What if it's violated? Violation is pretty serious because your model doesn't really make sense. First check if you didn't miss any terms in the models, perhaps a higher level effect is missing. If it's only a slight violation a Weighted Regression is a solution, which adds weights to scores and corrects for the "pull" of the high variance observations. Another population solution is to transform your data.

3.4.3 Multicollinearity

This one might still sound familiar from the multiple regression analysis and really only applies if you have more than one covariate in the model. The basic premise is that your variables don't correlate too much; if they do you are more or less measuring with the same stick. Imagine using both Height in Inch and Height in CM in your model, makes no sense, pick one!

How to check this? Easiest way to check is to run some correlations on your variables. There's no hard rule, but anything below 0.70 should be ok.

Correlations			
	Predictor 1	Predictor 2	Predictor 3
Predictor 1	1	.622	.330
Predictor 2	.622	1	.264
Predictor 3	.330	.264	1

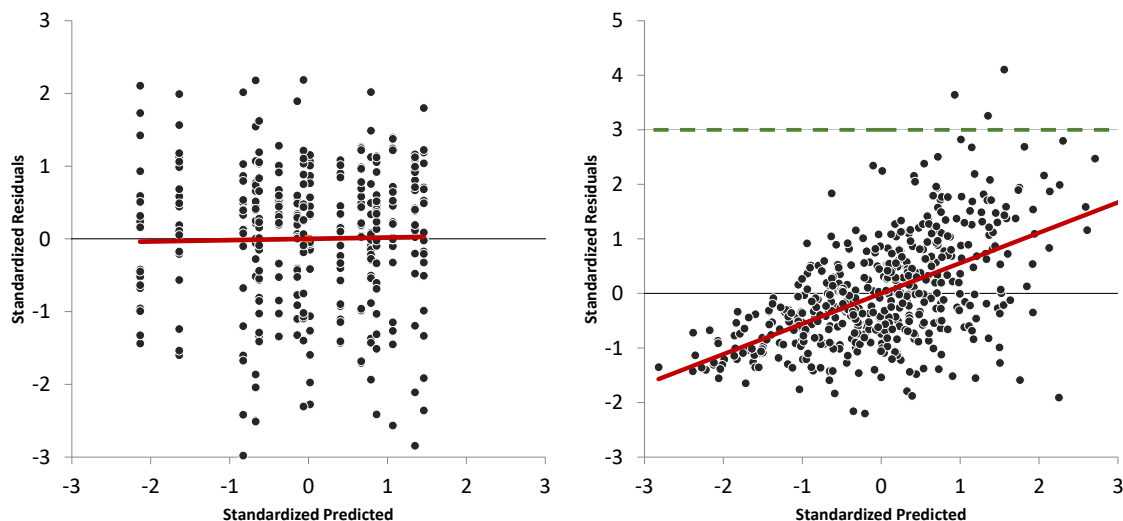
Correlations			
	Predictor 1	Predictor 2	Predictor 3
Predictor 1	1	.967	.456
Predictor 2	.967	1	.345
Predictor 3	.456	.345	1

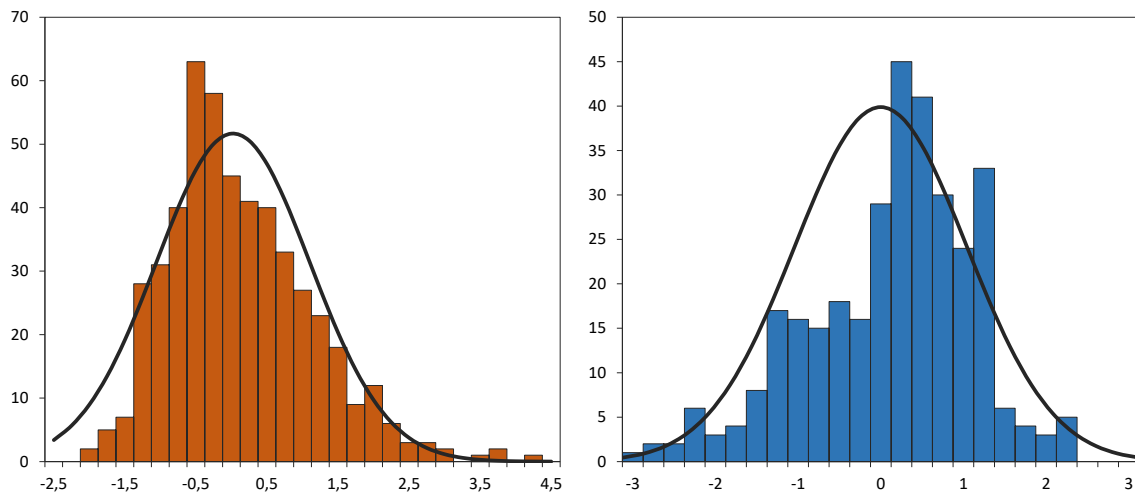
What if it's violated? Violation is easily solved by removing one of your variables. This is a bit of a theoretical exercise on what you should leave and how they are related to each other and the dependent. A more complicated alternative is to create a new variables using component- or factor-analysis.

3.4.4 Outliers and Normality

This is a familiar one, normality and outliers are pretty much important in every analysis. This has to do with the fact we use means and variances for much of the calculations so it is assumed they come from normal distributions.

How to check this? That scatterplot we made for homoscedasticity is multi-purpose! If you uses standardized predicted values residuals (Z-transform them) outliers are easy to spot. Anything outside of the -3 to 3 range can be considered an outlier. Normality can be checked using a histogram or Q-Q plot, the statistics work as well but those come out as non-normal more often than not.





What if it's violated? Outliers can be removed, not a big deal. Mixed Models excel in this aspect since you can remove individual observations and the model will run with whatever you have left. You might want to compare models to see how much the model changes after removing your outliers (one at a time!). If nothing changes you can consider leaving them in. Normality is a bit more flexible, most of the time models are robust against violations. You can try some data transformations (the log transform is most popular) and see how it affects the model. Other transformations are possible too, but they make interpretation of results quite difficult.

3.5 Building a Model

And here we are, the highlight of this session: How to build your Model. Building a model is a bit of an exercise where science meets art, it'll be different for everyone but I managed to get a 5-step plan that should work for most models. But before we do that we'll need to know how to compare two models, how to determine which model fits the data best.

3.5.1 Comparing two Models

Information Criteria^a

-2 Restricted Log Likelihood	2402,637
Akaike's Information Criterion (AIC)	2444,637
Hurvich and Tsai's Criterion (AICC)	2448,762
Bozdogan's Criterion (CAIC)	2539,249
Schwarz's Bayesian Criterion (BIC)	2518,249

We will be using the information criteria table for this part. The values shown are meaningless on their own (they scale with the data), but they are useful when comparing two similar models.

Many sources will tell you to compare the -2 Log Likelihood of a nested model to the -2LL of the larger model. The difference in -2LL and the difference in

parameters will act as a chi-square value and DF to determine if the increase is significant or not. This is a popular method but only works for nested models.

A nested model is a model that is a simplification of another model. The Compound Symmetry structure for example is a simplification of the Unstructured model and can be compared like this, but you cannot compare Compound Symmetry and AR(1), they are different but not nested.

Using BIC or AIC to compare two models usually protects against comparing non-nested models. These two values tend to agree, with the lowest value corresponding to the best fitting model. They are very similar in their likelihood transformation. $-2LL + kp$ where p is the number of parameters and k is 2 for AIC and $\log(n)$ for BIC.

- AIC is an estimate of a constant plus the relative distance between the (unknown) true likelihood and the fitted likelihood. Lower AIC means a model is closer to the "true fit".
- BIC is an estimate of a function of the posterior probability of a model being true. This is a Bayesian inference, meaning that a lower BIC is considered more likely to be true.

Both criteria are equally unrealistic yet correct, but in different ways. AIC has a chance of favoring the biggest model (with the most parameters) regardless of N . BIC on the other hand won't choose the bigger model if N is sufficient, but is more inclined to favor the smaller model. AIC and BIC are best used together. AIC works best when a false negative would be more misleading, while BIC is better in situations where a false positive is equally bad or worse than a false negative.

3.5.2 When is a model a better fit?

In practice you'll see that Unstructured will most often have the best fit, what we want to know is if we can sacrifice a little bit of accuracy for more degrees of freedom (and thus more power). I like to compare the AIC values. You find the lowest one (usually unstructured) and compare the 2nd lowest one to it. The difference between them determines whether or not the 2nd lowest is a good simplification.

<i>Difference between AIC_{min} and AIC_i</i>		<i>Interpretation</i>
<2 (36.79%)		Choose the i^{th} model, the evidence against this model is barely worth mentioning and the simplification will add degrees of freedom.
2-4 (13.53% - 36.79%)		There is strong support for the i^{th} model
4-7 (3.02% - 13.53%)		It's less likely that the i^{th} model is a good simplification
>10 (0.67%)		Stick with the AIC_{min} model, the loss of fit is not worth the simplification

If the AIC differences are a bit difficult to understand, the formula below transforms them into a percentage. The percentage can be read as the evidence for the i^{th} model being better. In the case of a 2-point difference this would be about 37%, which is quite high.

$$EXP\left(\frac{-(AIC_i - AIC_{min})}{2}\right) * 100\%$$

A last check that might be important in some cases is to see if a model is improving on actual fit instead of simply having different parameters. This is especially true for models with similar -2Log Likelihoods (because AIC will take parameters into account).

If $\frac{\Delta i}{2\Delta k} < 1$ then your model has a better fit and your AIC isn't just affected by the difference in the number of parameters.

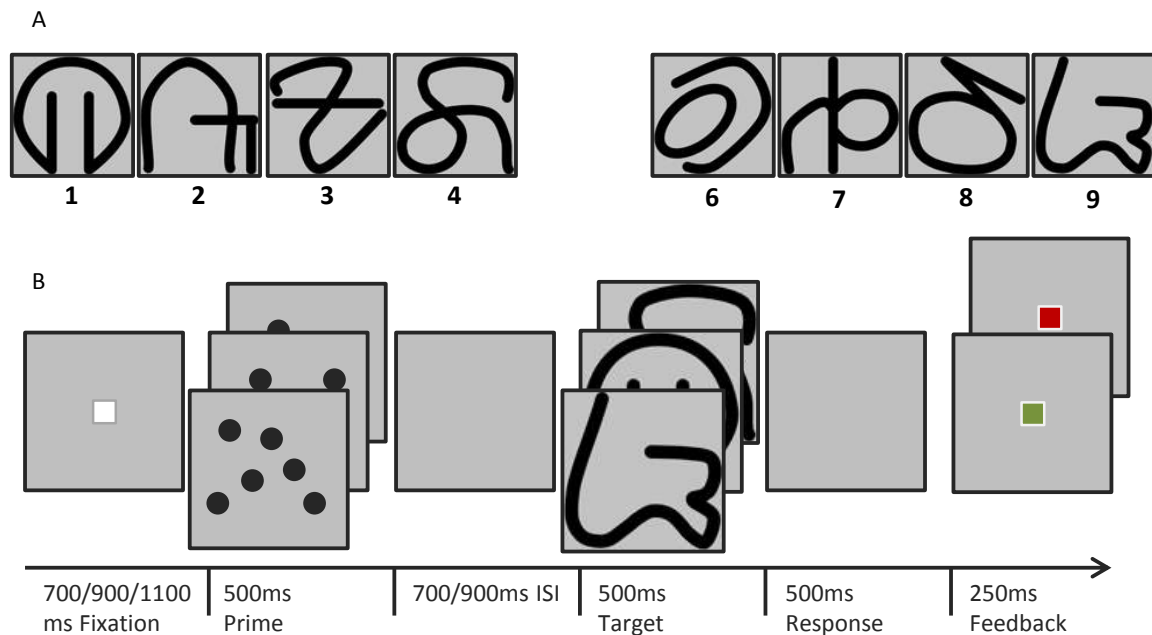
That's pretty much it; comparing AIC values will tell us which model fits best. We are now set up to start building models and see which one describes the data the most accurate.

3.5.3 Step 0: The Data

This is a pre-step, something not just for mixed models but for any model. **Know thy data!** Know where it's from, what you can expect from it, what the literature says. This will help you specify the best model.

The example we're working with is a simulation based on one of my own analyses. The experiment was a same-different choice task. Participants saw a prime in the form of a dot-pattern, then a target in the form of a symbol, and had to respond whether these represented the same quantity or not. They received feedback on their answer right away. This was done over three sessions, resulting in reinforced learning.

The data is simulated EEG data collected right after the Target. It has already been simplified by aggregating the data into groups, giving us **two Sessions (First and Last)**, **two Prime groups (Low and High)**, **two target groups (Low and High)**, and **two hemispheres (Left and Right)**, making this a 2x2x2x2 design.



3.5.4 Step 1: Factor Selection

Start with a full factorial model (if possible) and see if you need to transform any data. For now we prefer overfitting and we stick to **Maximum Likelihood (ML)** because we are adding or subtracting **fixed effects**. This is also a bit of a data-exploration phase. It might be better to combine certain variables into new factors or split a variable into multiples.

If you have repeated measures the covariance structure generally doesn't matter that much. I would suggest beginning with something simple like **compound symmetry** because that will run a lot faster and has a bigger chance of converging.

With this model we can also already check some of the assumptions that require residuals. Fun fact, your residuals don't change when you change covariance structures. We can already check for

outliers, normality, and homoscedasticity.

3.5.5 Step 2: Covariance structure selection

Now that the model contains all the factors and effects we want to add, we can start checking the covariance structure. First, **use your head!** Consider the various structures and think about which ones do and don't make sense. Is AR(1) a viable assumption or is Compound Symmetry more sensible? This will help you narrow down the options. My standard checklist is Unstructured, Toeplitz, Compound Symmetry, AR(1), and their heterogeneous versions.

Since we are changing the covariance matrix we are technically **changing a random factor**, so we need to switch to **Restricted Maximum Likelihood (REML)** to properly compare the models.

For my own peace of mind I usually summarize the Information Criteria in a table where it is easy to see how many parameters each model has and what the -2LL, AIC, and BIC are. Often AIC and BIC will agree, but with complex models they can differ.

They both suggest **AR1** or **Compound Symmetry** as the best fitting pattern, it even suggest the structure with equal variances as opposed to their heterogeneous counterparts. Now we have to think what is best, in this case I went for CSH because of the type of data we're dealing with. The assumption of equal covariances is fair, but assuming equal variances goes too far. AR1 would suggest a systematic increase with distance, but session was not equally spaced.

Structure	Parameters	-2LL	AIC	BIC
UN	152	519.580	791.580	1328.867
TPH	47	657.066	719.066	841.536
ARH(1)	33	667.526	701.526	768.687
CSH	33	667.780	701.780	768.941
TP	32	667.986	699.986	763.197
AR(1)	18	679.227	683.227	691.128
CS	18	679.392	683.392	691.293

3.5.6 Step 3: Model Reduction

With the covariance structure set up we can go back to the effects. We (probably) have way too many, all sorts of higher order interactions that are non-significant that we would like to get rid of. Less parameters means more degrees of freedom and removing a non-significant interaction can change the other effects for the better (especially in an unbalanced design, which you can get if you remove outliers). We switch back to **Maximum Likelihood (ML)** because we are messing with the fixed part of the model.

You start with the highest order effects, **4-way interactions** or higher are more often than not non-significant and not worth interpreting (explaining them is a bitch!). There are some studies that aim for such high interactions, but they will have studies specifically set up to detect these. The first thing you look at is significance, kicking out the least significant one first. A second criterion (definitely not less important) is theoretical meaning. Is there a reason to keep this effect in there? Do you expect an effect or is it important for you to make a point? If it is, then leave it in there. A final point to note is that you cannot remove an effect that is part of a higher order interaction. You need to keep a two-way interaction if you are also keeping its three-way interaction.

In the example below I removed the **4-way interaction**, then a **3-way interaction** that didn't include session (since session was most important from a theoretical background). The last one I removed was the **Session*Hemisphere*Prime** interaction because it wasn't that important from a theoretical point of view (no expectations) and it didn't change at all when removing the others.

The **Session*Prime*Target** was kept because that was an effect of interest, as was the **Hemisphere*Prime** interaction (although that one is debatable). All other 2-way interactions needed to be kept because they were **significant** or part of a **3-way interaction**.

Source	Sig.	Sig.	Sig.	Sig.
Intercept	.000	.000	.000	.000
Session	.000	.000	.000	.000
Hemisphere	.000	.000	.000	.000
Prime	.000	.000	.000	.000
Target	.000	.000	.000	.000
Session * Hemisphere	.671	.670	.651	.596
Session * Prime	.000	.000	.000	.000
Session * Target	.941	.954	.969	.956
Hemisphere * Prime	.796	.809	.816	.784
Hemisphere * Target	.065	.071	.065	.057
Prime * Target	.000	.000	.000	.000
Session * Hemisphere * Prime	.332	.338	.344	
Session * Hemisphere * Target	.000	.000	.000	.000
Session * Prime * Target	.053	.048	.047	.054
Hemisphere * Prime * Target	.676	.655		
Session * Hemisphere * Prime * Target	.531			

3.5.7 Step 4: The Final Model

We are finally done, we checked everything, removed all the unimportant effects and we can finalize the model. We're switching back to **Restricted Maximum Likelihood (REML)** because we have a mixed model (subject is random) and REML estimates are better for inference. Not much is going to change, all values are adjusted a bit because the DF changes.

This is the model you'll be using to interpret your data, so you might also want to order some post-hoc tests (we'll get into Estimated Marginal Means in Session 3). If you need peace of mind you can recheck the assumptions to make sure your final model still holds.

Type III Tests of Fixed Effects			
Source	DF	F	Sig.
Intercept	24.132	18995.535	.000
Session	343.185	70.955	.000
Hemisphere	343.756	63.420	.000
Prime	344.281	671.031	.000
Target	343.559	414.250	.000
Session * Hemisphere	342.950	0.272	.603
Session * Prime	344.428	186.437	.000
Session * Target	340.544	0.003	.956
Hemisphere * Prime	345.203	0.073	.788
Hemisphere * Target	343.152	3.531	.061
Prime * Target	344.165	19.956	.000
Session * Hemisphere * Target	347.960	17.249	.000
Session * Prime * Target	344.887	3.624	.058

3.5.8 Step 5: Writing it Down

We are finally done, we checked everything, removed all the unimportant effects and we can finalize the model. Writing down the process and results from a Mixed Model is a bit trickier than the ANOVA. It's still kind of a wild-west out there without standards. The APA has no official method of describing Mixed Models and most articles kind of do their own thing.

What is important is to **describe a model that can be understood**. Some journals/reviewers are more knowledgeable than others. As Mixed Models come into the mainstream this will be less of an issue and many researchers know how to interpret them, but better to be safe than sorry.

Tailor your description to the journal and audience. Highly technical articles for an audience familiar with Mixed Models can suffice by mentioning the factors and the covariance structure. Others might need more on why you are using Mixed Models, they might be suspicious that you're messing around with complex models to get results you want.

With the lack of direction the best way to know what to report in your article is to look at other articles in your area. Keep in mind that your analysis must be reproducible from your description, so keep in mind a few questions:

- What were the Fixed Effects in the model?
 - *Were they centered, coded, scaled?*
- What was the covariance structure?
- How were the Fixed effect and Covariance structure determined?
- How were different models evaluated?

- What software was used?
- Which method of estimation was used (REML/ML)?
- Were assumptions violated and/or corrected?
- Was there missing data (did it impact the model)?

For this example the piece below might be a good start, depending on the journal it is being sent to.

Statistical Analysis

A Linear Mixed Model was constructed (IBM SPSS 24) using within-subject factors Session (2 Levels; Session 1 & Session 3), Hemisphere (2 Levels; Left & Right), Prime (2 Levels; Low & High), and Target (2 Levels; Low & High). Covariance Structures were compared on Restricted Maximum Likelihood (REML) models using Akaike's Information Criterion (AIC).


Non-significant higher order interactions were removed on Maximum Likelihood (ML) models in a step-wise manner starting with the four-way interaction Session*Hemisphere*Prime*Target ($p=.531$), followed by the three-way interaction between Hemisphere*Prime*Target ($p=.655$) and Session*Hemisphere*Prime ($p=.344$). The Session*Prime*Target interaction was kept based on theoretical expectations and all two-way interactions also remained in the model.

The final model used for the analyses (reported below) was a fixed effects model (REML) using a Heterogeneous Compound Symmetry Covariance Matrix for repeated measures. Residuals were normally distributed, showed no heteroscedasticity, and no observations were removed as outliers.

Results

The Linear Mixed Procedure on [DATA] showed main effects for Session ($F_{(1,343.18)}=70.955$, $p<.001$), Hemisphere ($F_{(1,343.756)}=63.42$, $p<.001$), Prime ($F_{(1,344.28)}=671.03$, $p<.001$), and Target ($F_{(1,343.56)}=414.25$, $p<.001$).

There were two-way interactions of Session*Prime ($F_{(1,344.43)}=186.44$, $p<.001$) and Prime*Target ($F_{(1,344.17)}=19.96$, $p<.001$). Lastly there was a three-way interaction between Session, Target, and Hemisphere ($F_{(1,347.96)}=17.25$, $p<.001$).



4. Extending the Model

Introduction

In the previous sessions we worked on building models, what to look at to make sure your final model has the best fit. With a nice model set to go we can continue to mess it up again, or more nicely put "extending" it. By extending I mean looking at the effects we detected as significant, looking at the main effects and interactions.

Instead of running more models per levels of some other variable to get parameters that only compare two groups, we'll save ourselves some time by ordering **Estimated Marginal Means**. If done right you will end up with a single model that contains all pairwise comparisons you need to draw conclusions.

Then we ramp things up and look at the inclusion of **covariates**, which is the term for continuous variables. These variables are a bit different than factors since they don't have levels, they have values. Adding covariates can be very useful though, and we even look at a case of **moderation** where the value of a covariate changes the effect of a factor. By using **Estimated Marginal Means** it is possible to also order everything you need for a **moderation analysis** in a single model.

The second section of this sessions covers **Random Effects**. This includes the includes what they are, adding random intercepts, and adding random slopes.

Last time (recap)

In the last analysis we went through the steps of finding the right matrix and reducing the model. We decided on a **Fixed Effects model** with a **Heterogeneous Compound Symmetry Structure** and we removed the 4-way interaction and some of the 3-way interactions. But as you can imagine, this is not the end.

Type III Tests of Fixed Effects			
Source	DF	F	Sig.
Intercept	24.331	241.283	.000
Session	150.949	25.635	.000
Hemisphere	283.510	25.779	.000
Prime	303.339	228.323	.000
Target	293.658	247.809	.000
Session * Hemisphere	284.574	0.098	.754
Session * Prime	301.728	88.368	.000
Session * Target	300.697	2.731	.099
Hemisphere * Prime	301.215	0.656	.419
Hemisphere * Target	308.144	1.803	.180
Prime * Target	268.941	5.570	.019
Session * Hemisphere * Target	297.199	5.054	.025
Session * Prime * Target	288.018	0.001	.981

We are left with some effects of interest, mainly the Session*Prime effect, but also a Prime*Target effect and a suspicious Session*Target*Hemisphere effect. What are we going to do? Split-file and check the effect of Prime per Session? That is a possibility but probably a better idea for the three-way interaction. The three-way interaction is counterintuitive in this design, and I would suspect a false positive there. Splitting the data and looking at the interactions would inform us if there really is something there (checking the Hemisphere*Target interaction per session would be the most interesting). For the two-way interactions (and higher interactions that aren't that suspicious) a better option is to look at Estimated Marginal Means. Why run more models on split data if you can simply order the means and pairwise comparisons you need?

4.1 Post-Hoc Testing (Estimated Marginal Means)

I know it seems like a super-obvious thing, but I'm always surprised at how often this doesn't happen. The reason for this oversight: most statistics programs (like SPSS) won't automatically give you EMMs for interactions. The computer doesn't know what you want to compare, and it won't give you all possible combinations because that can get very processor heavy in large models. The solution is for us to tell the computer what we want to compare. In SPSS we will need to adjust the Syntax a little bit, in other languages similar commands are needed. It's as simple as adding **COMPARE (Session)** to the EMM command and specify the multiple testing correction using **ADJ (SIDAK)**.

```
/EMMEANS=TABLES (Session) COMPARE ADJ (SIDAK)  
/EMMEANS=TABLES (Session*Prime) COMPARE (Session) ADJ (SIDAK)
```

With that SPSS knows what we want to compare and provides us with **Pairwise Comparisons** we can use to determine where the interactions were coming from.

4.1.1 What are EMMs?

Remember that regression models give you all those coefficients? Those beta values? That's pretty much your model. The Intercept is the default state, the reference category. Coefficients are also the same as in normal regression, it is the linear change (increase or decrease) of the predicted value when you add 1 to this effect. It doesn't matter that we are using categories, the concept is still the same.

When we added these predictors as factors (indicating that they are categorical) every unique value was transformed into a dummy variable. If you have three values (1, 2, 3) you get two dummy variables: [Value=1] and [Value=2]. The last category is the reference category (at least in the SPSS Mixed Procedure, other programs like R use the first), which means it is included in the Intercept (if you're not in 1 or 2 you are in category 3). If you ever mess up and add a continuous predictor as a factor instead of a covariate, you will see a lot of different predictors show up (one for every unique value).

In the model the Intercept represents **[Session=3]-[Hemisphere=1]-[Prime=1]-[Target=1]** and the predicted value is -2.099, that's the base, the reference. X1 is the **[Session=1] Dummy Predictor** and b1 is the **[Session=1] Effect** of changing the dummy from 0 to 1. The table tells us that the difference between the **reference** and **[Session=1]** is 0.088, in other words: we add 0.088 to the prediction if we switch from Session 3 to Session 1.

Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	Sig.
Intercept	-2.099334 (β_0)	.095524	34.630	-21.977	.000
X_1 - [Session=1]	0.088284 (β_1)	.138783	83.107	0.636	.526
X_2 - [Hemisphere=0]	-0.268412 (β_2)	.126612	113.425	-2.120	.036
X_3 - [Prime=0]	-1.777300 (β_3)	.118407	89.862	-15.010	.000
X_4 - [Target=0]	-0.643985 (β_4)	.144279	89.972	-4.463	.000
X_5 - [Session=1]*[Hemisphere=0]	-0.511187 (β_5)	.156703	179.406	-3.262	.001
X_6 - [Session=1]*[Prime=0]	1.266037 (β_6)	.155710	189.037	8.131	.000
X_7 - [Session=1]*[Target=0]	-0.668476 (β_7)	.197265	159.145	-3.389	.001
X_8 - [Hemisphere=0]*[Prime=0]	-0.029592 (β_8)	.109820	345.203	-0.269	.788
X_9 - [Hemisphere=0]*[Target=0]	-0.248907 (β_9)	.157123	175.478	-1.584	.115
X_{10} - [Prime=0]*[Target=0]	-0.691462 (β_{10})	.156733	186.827	-4.412	.000
X_{11} - [Session=1]*[Hemisphere=0]*[Target=0]	0.908256 (β_{11})	.218691	347.960	4.153	.000
X_{12} - [Session=1]*[Prime=0]*[Target=0]	0.416681 (β_{12})	.218891	344.887	1.904	.058

With these coefficients we can build a regression formula. This is how the model predicts what the value is going to be, multiplying each coefficient by the values of the predictors. In this example it's very easy, we only have factors, only have dummy variables, so the predictors can only be 0 or 1 (either you're in that category or not). This makes calculating the predicted value a matter of summing up the relevant coefficients. Later we will look at covariates, which is a little bit more complicated but basically the same thing (instead of zero or one it takes on a value).

$$\hat{Y} = \beta_0 + (X_1\beta_1) + (X_2\beta_2) + (X_3\beta_3) + (X_4\beta_4) + (X_5\beta_5) + (X_6\beta_6) + (X_7\beta_7) + (X_8\beta_8) + (X_9\beta_9) + (X_{10}\beta_{10}) + (X_{11}\beta_{11}) + (X_{12}\beta_{12})$$

We can fill in the coefficients for each beta and end up with the formula below:

$$\hat{Y} = -2.099 + (X_1 0.088) + (X_2 - 0.268) + (X_3 - 1.777) + (X_4 - 0.644) + (X_5 - 0.511) + (X_6 1.266) + (X_7 - 0.668) + (X_8 - 0.029) + (X_9 - 0.249) + (X_{10} - 0.691) + (X_{11} 0.908) + (X_{12} 0.417)$$

That's all there is to it, that is what EMMs are. They are the results, the predicted values, for a given combination of predictors. In the current example we can have 16 different predictions ($2 \times 2 \times 2 \times 2 = 16$) and we can write them all out. Pairwise comparisons are then applied to these predictions, which in the end is what we want to compare and why we want a good fitting model. When we order the Session*Prime interaction for example, it gives us a prediction for all four level of this interaction, averaging over the other predictors. This estimation is the same as the average of the predictions below, though the computer uses a more direct way to calculate it (in this case it sets the Hemisphere and Target effects to 0.5).

Session	Hemisphere	Prime	Target	Prediction	Session	Prime	Prediction
1	Left	Low	Low	-4.259	1	Low	-3,556
1	Right	Low	Low	-4.110	1	High	-2,892
1	Left	High	Low	-3.444	3	Low	-4,756
1	Right	High	Low	-3.324	3	High	-2,618
1	Left	Low	High	-3.332			
1	Right	Low	High	-2.522			
1	Left	High	High	-2.791			
1	Right	High	High	-2.011			
3	Left	Low	Low	-5.759			
3	Right	Low	Low	-5.212			
3	Left	High	Low	-3.261			
3	Right	High	Low	-2.743			
3	Left	Low	High	-4.175			
3	Right	Low	High	-3.877			
3	Left	High	High	-2.368			
3	Right	High	High	-2.099			

Estimated Marginal Means are an incredible useful tool and they represent the effects we want to investigate. They rely on the model, the better the model the better the estimation, which makes it even more important to ensure a good fit. Pairwise comparisons are then applied to the estimations, giving you the significance values to report in your results (don't forget to adjust for multiple comparisons).

4.2 Adding Covariates to the Model

Up to this point we were using a 2x2x2x2 repeated measures design and all the predictors we put in there were factors. While easier designs are wanted and preferred, we aren't always that lucky and simply need more complicated designs. A big advantage of the Mixed Procedure, and the Univariate data structure it uses, is the addition of covariates. During many basic Statistics courses Covariates are introduced with the ANCOVA model. In these models a covariate is included to control for confounding, a common reason to use them. While not wrong it makes it seem like a covariate is "a predictor we control for", which is a bit misleading. A Covariate is nothing more than a continuous predictor, with Factors being categorical predictors. If the covariate is significant but doesn't interact with anything else, you have a model similar to the ANCOVA where you control for confounding. If the covariate also interacts with some of the other predictors it's a moderation model, something we'll see later.

The example we'll be working with here is the same experiment as before, but with different data to show how to include a covariate, how controlling for a covariate can have serious effects on your conclusions, and how to deal with moderation.

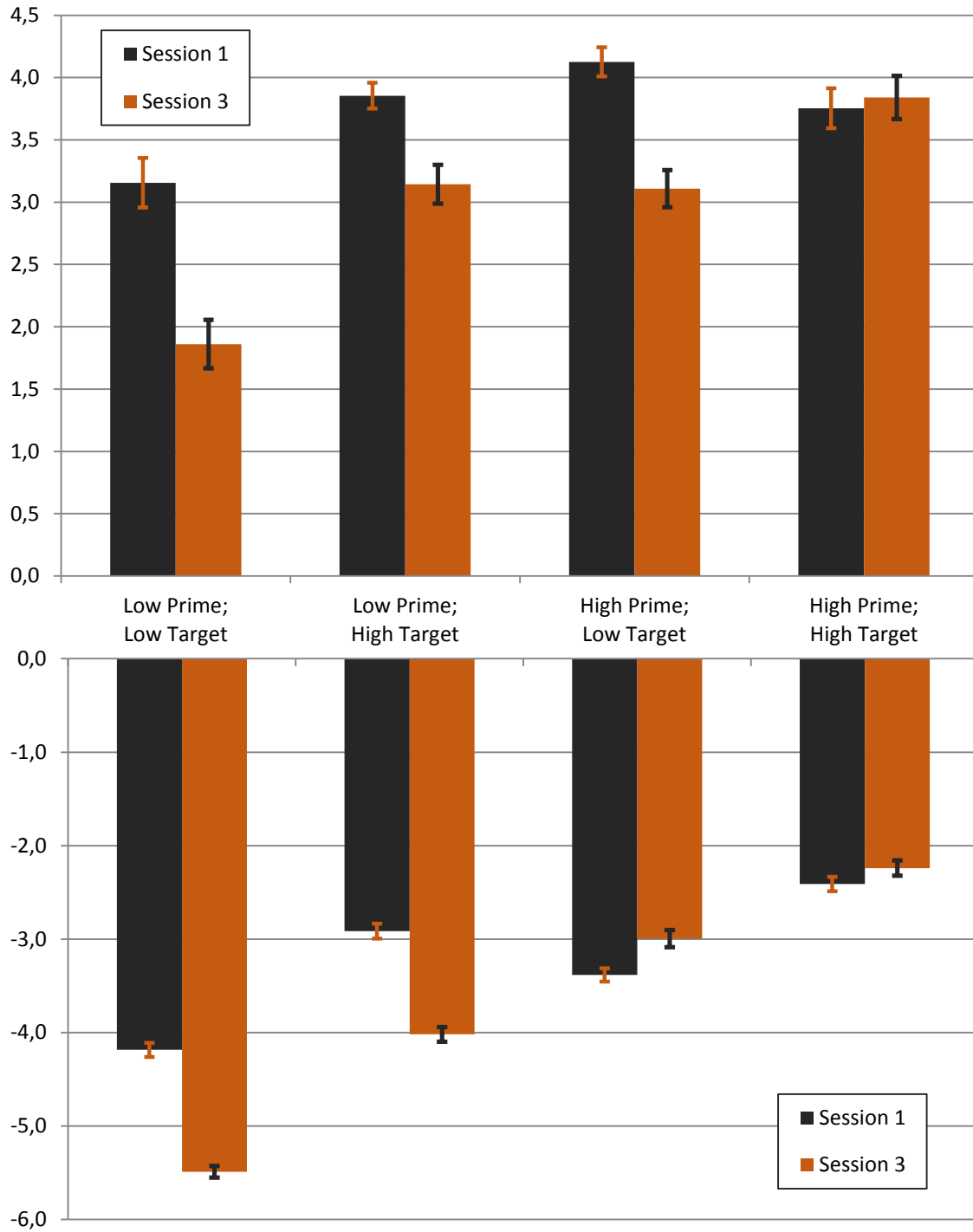
The previous model was applied to a visual EEG component, a neural response that comes quickly after the stimulus presentation ($\pm 100\text{ms}$). The current data comes from a subsequent response, an internal process that focuses on the numerical meaning of the stimulus ($\pm 220\text{ms}$). We'll first look at the **Numerical Response** without taking the **Visual Response** into account. After that we will include the **Visual Response as a covariate** to see what happens.

4.2.1 Without Covariate

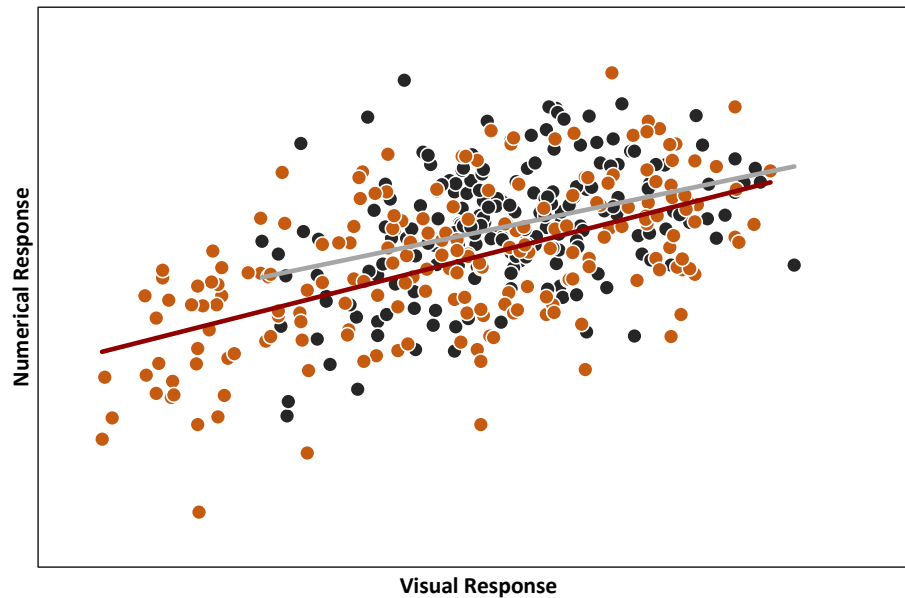
I went ahead and did all the steps for this response, finding the covariance matrix, and reducing the model. There are a lot of two-way interactions while the three-way interaction of interest (Session*Prime*Target) is not significant ($p=.211$). That's a bit weird, we kind of expect and effect here. You could continue, test the interactions, and report the results, but let's investigate.

Type III Tests of Fixed Effects										
Source		Sig.	Sig.	Sig.	Sig.	Sig.	Before	Source		
	Intercept	.000	.000	.000	.000	.000	.000	Intercept		
	Session	.000	.000	.000	.000	.000	.000	Session		
	Hemisphere	.058	.072	.134	.279	.272	.000	Hemisphere		
	Prime	.121	.128	.126	.118	.159	.000	Prime		
	Target	.306	.319	.296	.238	.330	.000	Target		
	Session * Hemisphere	.606	.518	.478	.265	.298	.107	Session * Hemisphere		
	Session * Prime	.065	.068	.064	.060	.125	.011	Session * Prime		
	Session * Target	.000	.000	.000	.000	.000	.000	Session * Target		
	Hemisphere * Prime	.003	.004	.007	.014	.011	.455	Hemisphere * Prime		
	Hemisphere * Target	.727	.725	.969	.654	.567	.048	Hemisphere * Target		
	Prime * Target	.017	.020	.020	.022	.016	.000	Prime * Target		
	Session * Hemisphere * Prime	.180	.196	.272				Session * Hemisphere * Prime		
	Session * Hemisphere * Target	.061	.087	.059	.070			Session * Hemisphere * Target		
	Session * Prime * Target	.034	.036	.031	.033	.034	.211	Session * Prime * Target		
	Hemisphere * Prime * Target	.110	.137					Hemisphere * Prime * Target		
	Session * Hemisphere * Prime * Target	.238						Session * Hemisphere * Prime * Target		
	Visual_Component	.000	.000	.000	.000	.000		Visual_Component		
	Visual_Component * Hemisphere	.002	.002	.005	.010	.008		Visual_Component * Hemisphere		

Looking at the estimates, you might notice what is going on. The earlier **Visual Response** is dragging down the later **Numerical Response**, but only in the two Prime*Target categories on the left. That's probably confounding, meaning that the model we just ran on the **Numerical Response** also include effects on the **Visual Response**.



Would including the Visual Response as a predictor for the Numerical Response be a good idea? In more general terms we're asking if there is a correlation between the value we want to predict and some other continuous data that we want to include in the model. For our example, yeah looks like it's a good idea to include it. There a significant correlation between them, overall and for the separate sessions.



4.2.2 With Covariate

Now to do it all again, but including an extra predictor: the covariate. The results look a bit... scary, but it really isn't. Remember, before we went up to a 4-way interaction (16 effects in total), this time around we go up to a 5-way interaction. The Visual Response adds a single effect and it then interacts with all the other predictors and interactions, adding 16 more effects.

Type III Tests of Fixed Effects			
Source	Denominator df	F	Sig.
Intercept	368.000	251.145	.000
Session	366.527	.309	.579
Hemisphere	361.385	2.242	.135
Prime	359.931	1.545	.215
Target	363.405	4.668	.031
Session * Hemisphere	365.265	.000	.986
Session * Prime	367.845	1.953	.163
Session * Target	367.620	.143	.706
Hemisphere * Prime	366.721	.083	.774
Hemisphere * Target	366.152	.082	.774
Prime * Target	362.653	2.613	.107
Session * Hemisphere * Prime	362.349	.431	.512
Session * Hemisphere * Target	367.194	.318	.573
Session * Prime * Target	364.155	.267	.605
Hemisphere * Prime * Target	364.909	.239	.625
Session * Hemisphere * Prime * Target	367.777	.043	.836
Visual_Response	363.676	45.150	.000
Session * Visual_Response	367.001	.091	.764
Hemisphere * Visual_Response	361.058	8.404	.004
Prime * Visual_Response	361.507	.471	.493
Target * Visual_Response	364.146	3.456	.064
Session * Hemisphere * Visual_Response	365.626	.028	.867
Session * Prime * Visual_Response	367.975	1.475	.225
Session * Target * Visual_Response	367.907	.006	.940
Hemisphere * Prime * Visual_Response	367.615	.227	.634
Hemisphere * Target * Visual_Response	364.356	.308	.579
Prime * Target * Visual_Response	365.326	2.608	.107
Session * Hemisphere * Prime * Visual_Response	363.273	.111	.739
Session * Hemisphere * Target * Visual_Response	366.379	.921	.338
Session * Prime * Target * Visual_Response	363.339	.047	.828
Hemisphere * Prime * Target * Visual_Response	362.762	.185	.668
Session * Hemisphere * Prime * Target * Visual_Response	367.786	.060	.806

I won't make you suffer going through model reduction for this one, I'll just give you the answer. In this model the **Visual Response** is a significant predictor for the **Numeric Response**, but it also has an interaction with **Hemisphere**. The rest of the terms aren't significant and have been removed. After that I went through the same steps with the regular effects, removing the **4-way interaction** and some of the **3-way interactions**.

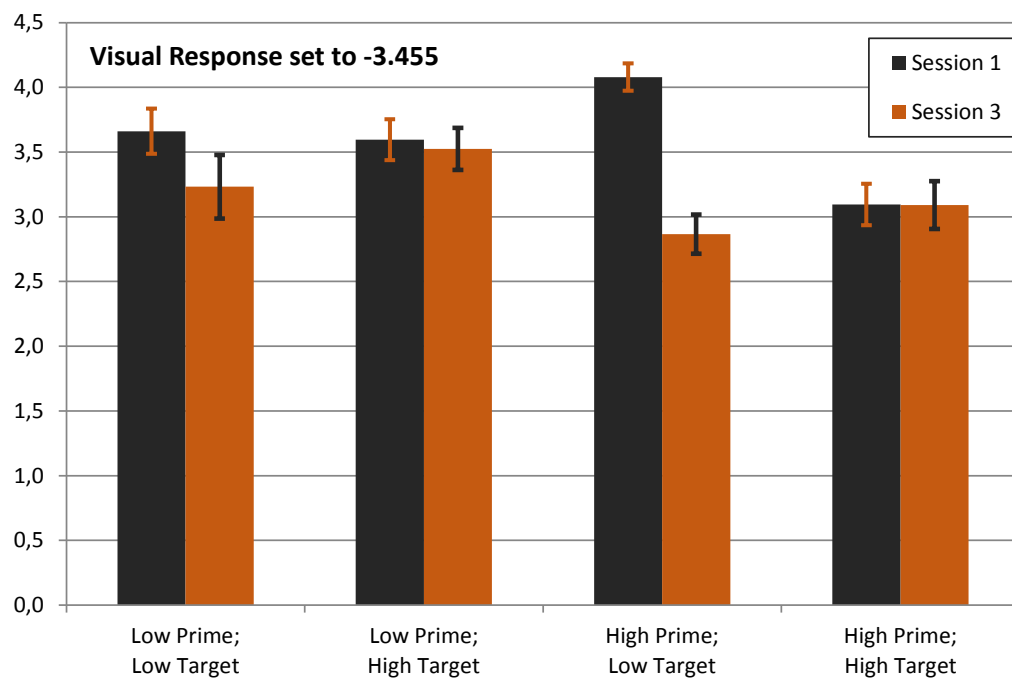
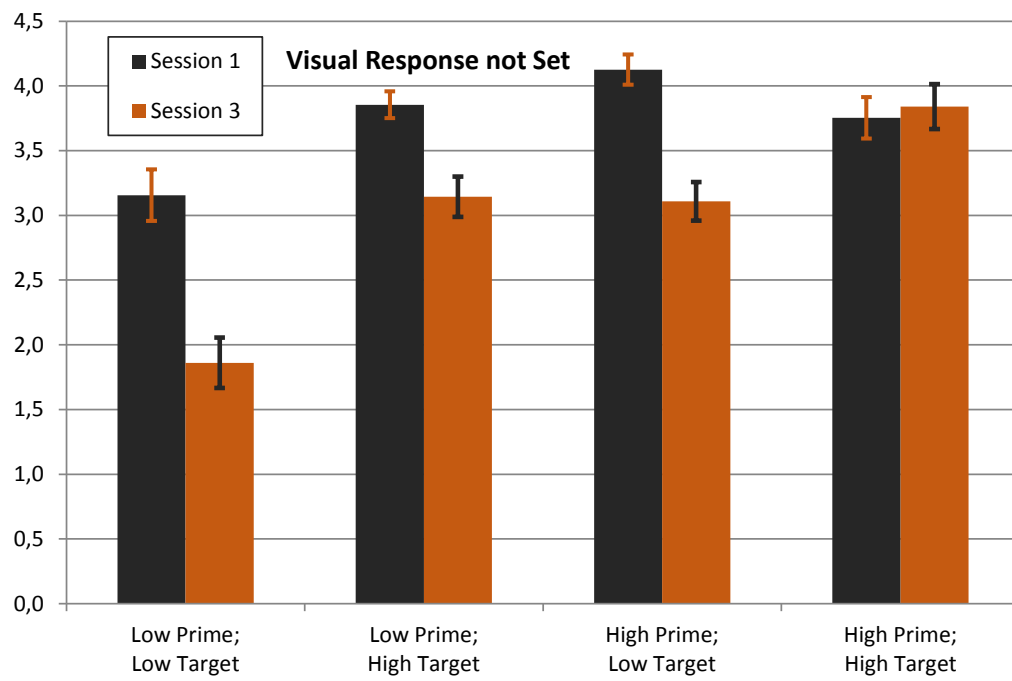
The inclusion of **Visual Response** as a covariate had some dramatic effects. Various effects have disappeared, becoming non-significant. The three-way interaction of **Session*Prime*Target** is now significant.

Type III Tests of Fixed Effects						Before	Source
Source	Sig.	Sig.	Sig.	Sig.	Sig.		
Intercept	.000	.000	.000	.000	.000	.000	Intercept
Session	.000	.000	.000	.000	.000	.000	Session
Hemisphere	.058	.072	.134	.279	.272	.000	Hemisphere
Prime	.121	.128	.126	.118	.159	.000	Prime
Target	.306	.319	.296	.238	.330	.000	Target
Session * Hemisphere	.606	.518	.478	.265	.298	.107	Session * Hemisphere
Session * Prime	.065	.068	.064	.060	.125	.011	Session * Prime
Session * Target	.000	.000	.000	.000	.000	.000	Session * Target
Hemisphere * Prime	.003	.004	.007	.014	.011	.455	Hemisphere * Prime
Hemisphere * Target	.727	.725	.969	.654	.567	.048	Hemisphere * Target
Prime * Target	.017	.020	.020	.022	.016	.000	Prime * Target
Session * Hemisphere * Prime	.180	.196	.272				Session * Hemisphere * Prime
Session * Hemisphere * Target	.061	.087	.059	.070			Session * Hemisphere * Target
Session * Prime * Target	.034	.036	.031	.033	.034	.211	Session * Prime * Target
Hemisphere * Prime * Target	.110	.137					Hemisphere * Prime * Target
Session * Hemisphere * Prime * Target	.238						Session * Hemisphere * Prime * Target
Visual_Component	.000	.000	.000	.000	.000		Visual_Component
Visual_Component * Hemisphere	.002	.002	.005	.010	.008		Visual_Component * Hemisphere

The significance values for the Fixed Effects table are based on a +1 change in the value of the effect, whilst all others are kept at 0. This was fine before because it meant we were comparing it to the reference category, but with a covariate we need to consider the value of 0. Is zero a meaningful value? For some variables it is whilst for others it makes no sense (like weight or height). It's a good idea to center your values by subtracting the mean, this changes the mean to 0. With the mean set to 0 all the effects in the Fixed Effects table represent the effects if the covariate is average.

The **Estimated Marginal Means** have also changed. They include a small subscript, indicating the value of the **Covariates**. By default, this value is set to the average of the covariate for computing the EMMs.

It seems like including the **Visual Response** was a good idea. Before three out of four Prime*Target combinations showed session effects. The two on the left seemed to be driven by the **Visual Response** rather than the **Numerical Response**. Setting the **Visual Response** to the mean (-3.455) removed this influence, giving us new estimated. These new estimated are basically: *If the visual response was -3.455 then the Numerical response is Y*. Lo and behold, the two session effects disappear and we are left with only one, drastically changing the interpretation and conclusion.



4.2.3 Dealing with Moderation

The fun isn't over yet, the model also included that Visual Response*Hemisphere interaction. This interaction means that the effect of **Visual Response** on **Numerical Response** is different in the Left and Right Hemisphere. We're not really interested in the **Visual Response** effect though, instead we're more curious about the **Hemisphere** effect, which is different for different values of the **Visual Response**.

Investigating a Continuous*Categorical interaction is tricky. For those who still remember their statistics course, post-hoc analysis on a **moderation model** involves running multiple models. This would mean that we make several **Visual Response** variables: *Low*, *Average*, and *High*.

The general advice is to add and subtract the **standard deviation** from the **centered value**. If you subtract 1SD then the average becomes +1SD (you shift the 0 to -1 and +1 becomes 0). Putting that in the model instead of the centered variable shows you the effect *if the moderator (visual response for us) is high*. Adding 1SD does the opposite and gives us the effect of a *Low Visual Response*.

This works, but it's rather laborious and for many confusing (what is high and what is low?). We are not going to bother with running multiple models, we don't have time for that and we're way to cool to go for such standard methods. Instead we will modify our current model to show us these values instead.

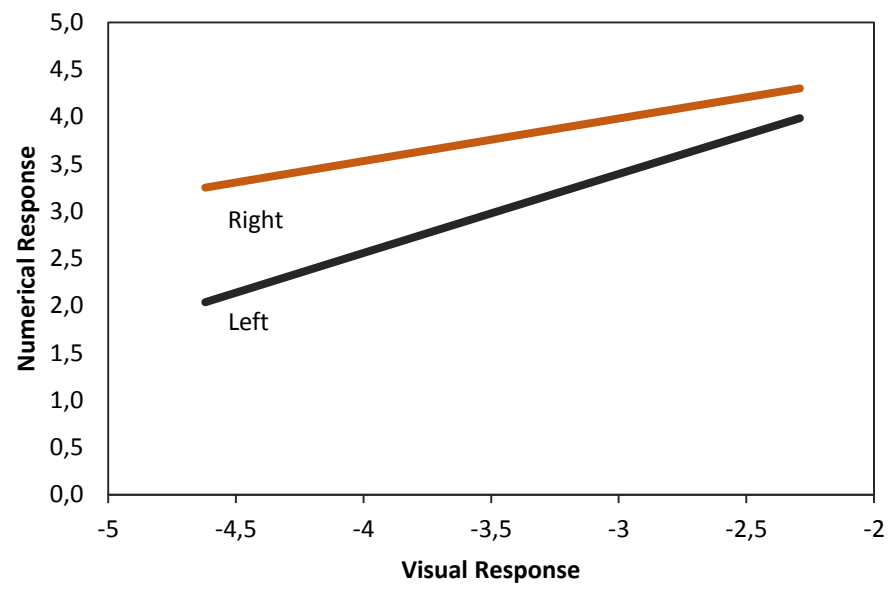
EMMs in moderation analyses are awesome, allowing you to specify the value for the moderator. In SPSS the EMM command can be extended using **WITH ([Moderator]=[Value])**. We can specify any value but I'd suggest the mean and the values corresponding to -1 and +1 SD. If you have centered your variable this corresponds to 0; -SD, SD (or even 0; -1, and 1 if you standardize).

```
/EMMEANS=TABLES (Hemisphere) WITH (VisualResponse=[-1SD]) COMPARE ADJ (SIDAK)
/EMMEANS=TABLES (Hemisphere) WITH (VisualResponse=[Mean]) COMPARE ADJ (SIDAK)
/EMMEANS=TABLES (Hemisphere) WITH (VisualResponse=[+1SD]) COMPARE ADJ (SIDAK)
```

This will prompt SPSS to provide you with three EMM tables for the **Hemisphere** effect, one for each value of the **Visual Component** we specified (the exact value is shown in the table notes).

For completeness we can see that *the difference between hemispheres is larger if the Visual Response was strong (very negative) and shrinks as the response lessens (becomes more positive)*.

Estimates 1SD Below				Estimates Mean				Estimates 1SD Above			
Hemisphere	Mean	SE	df	Hemisphere	Mean	SE	df	Hemisphere	Mean	SE	df
Left	2,037	,136	156,210	Left	3,011	,082	68,360	Left	3,985	,181	179,276
Right	3,250	,171	196,271	Right	3,775	,082	63,750	Right	4,300	,133	144,265
Visual Response = -4.62				Visual Response = -3.46				Visual Response = -2.29			



4.3 Random Effect Models

It is that time again, the time to ramp things up another notch. A mixed model technically is a model that contains both Random and Fixed effects. We've sort of used Mixed Models up until now because we had that implicit random **Person** effect in it.

Random effect models are probably less likely to occur for Cognitive Neuroscience experiments, you more often see them in Multi-level analyses. I have used them before when the repeated measure (like session) differs for each participant. This is more likely in learning studies, where different individuals learn at different rates.

A multi-level analysis has, well, multiple levels. **Person** is one level, within person the other factors vary. But this person can also be inside a classroom, in a school, in a city, in a county. Within each of these levels (class, school, city, country) you can expect that people within the group correlate more than people between groups (sound familiar?).

We can expect that children in the same classroom will score more similar than children in different classrooms (common effects such as having the same teacher). You can expect patients treated by the same physician to correlate more than patients treated by different physicians, cells from the same dish to be more similar than cells from different dishes, you get the idea.

One final, but important, example is something you might run into if you ever do a meta-analysis. The data from people within the same study will show more similar effects than the data from two different studies. It could be due to the studies taking place in different countries or due slight changes to the methods. Nevertheless, the effects reported are all the same effect, some kind of learning effect, treatment effect, or phenomena that we assume is the same.

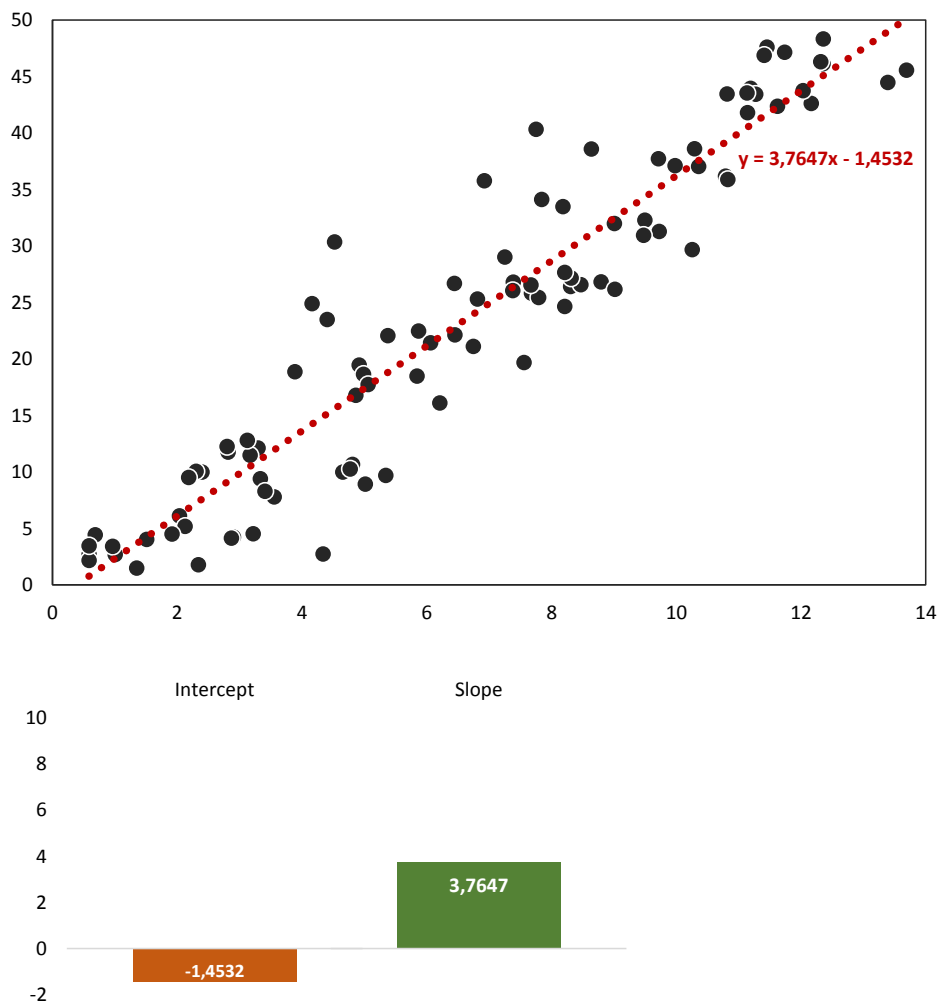
How will we take this difference into account? First instinct might be to include a between subject variable (such as petri-dish, classroom, or study). The idea is sound, if there is no difference between the various studies in the meta-analysis you might find a main effect of Study, but the Study*Effect interaction is not significant. The effect of interest can then be investigated and conclusions can be made.

Sounds good in theory, but in practice you'd be making the wrong assumptions. Such a model that includes a Between-Subjects effects assumes that both the Study and Effect we're looking at are fixed and independent.

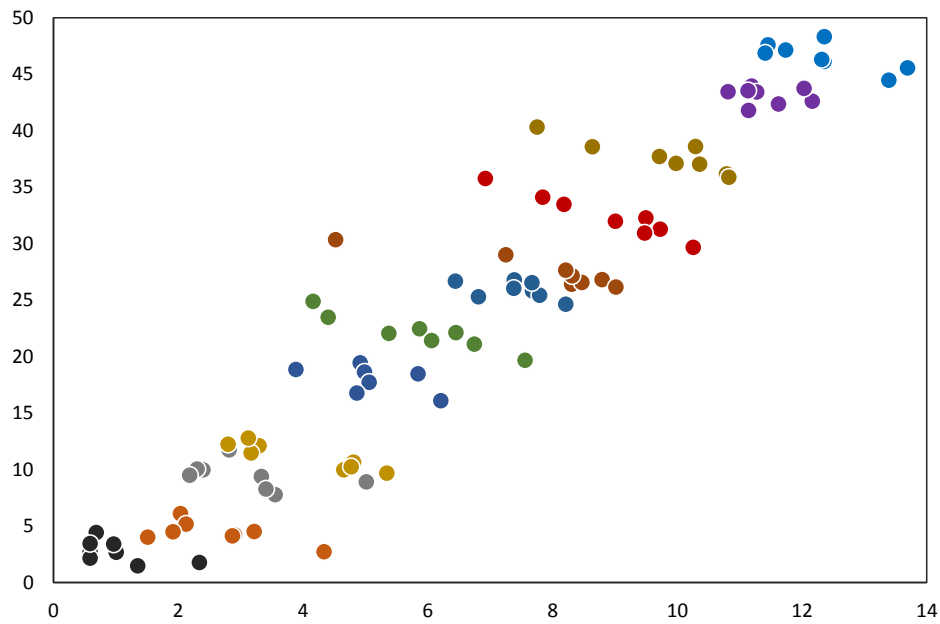
That's not what's going on. The various studies in a meta-analysis are similar to the Person effect, they are a (supposed) random sample from the population of studies and have dependency. We don't care about the difference between Study 1 and Study 2, just about a general Study effect. Subject within a study are probably more similar in how they respond; a between subject model won't take this dependency into account. If this is still a bit cryptic, don't worry. The example hopefully makes things clearer.

4.3.1 Model 1: Fixed Effects Model

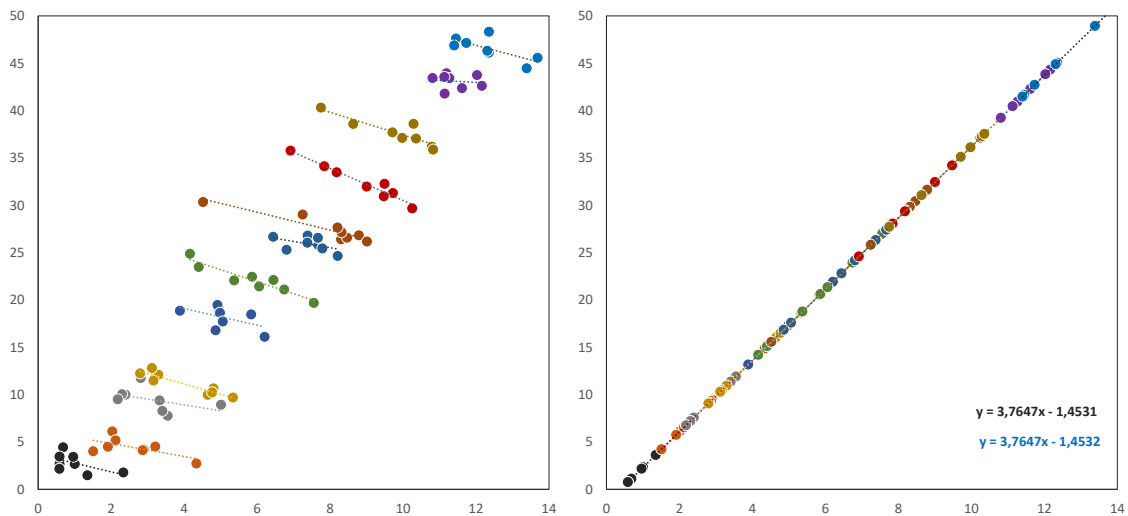
We're keeping things easy for now, no complicated 2x2x2x2 repeated measures design. Instead a simple single-predictor model where X predicts Y. The analysis seems pretty straightforward, the intercept is -1.45 and for every increase of X we can see that Y increases by 3.76 (a positive effect). Both of these effects are fixed. We assume that everyone in our sample has the same Intercept and Slope, any deviation is due to error; events outside the scope of the experiment.



But guess what? The example is a multi-level dataset. It contains data from 12 different studies (or hospitals, or dishes). You might see the problem, if we ignore the subgroups the effect is clearly positive, but within each subgroup the effect is negative. Going for model 1 would be a huge mistake, leading to the opposite conclusion.



Comparing the predicted values to the Observed values shows the problem more clearly. Oof, that's a bad prediction. The model thinks that the deviation from the linear fit is error, while in reality it's caused by the subgroups. Let's try something else.

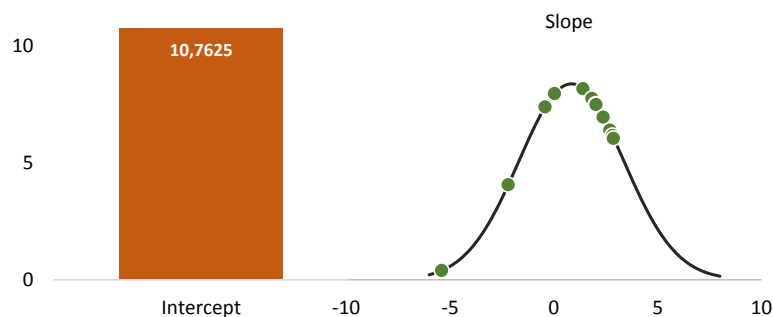


4.3.2 Model 2: Random Slope Model

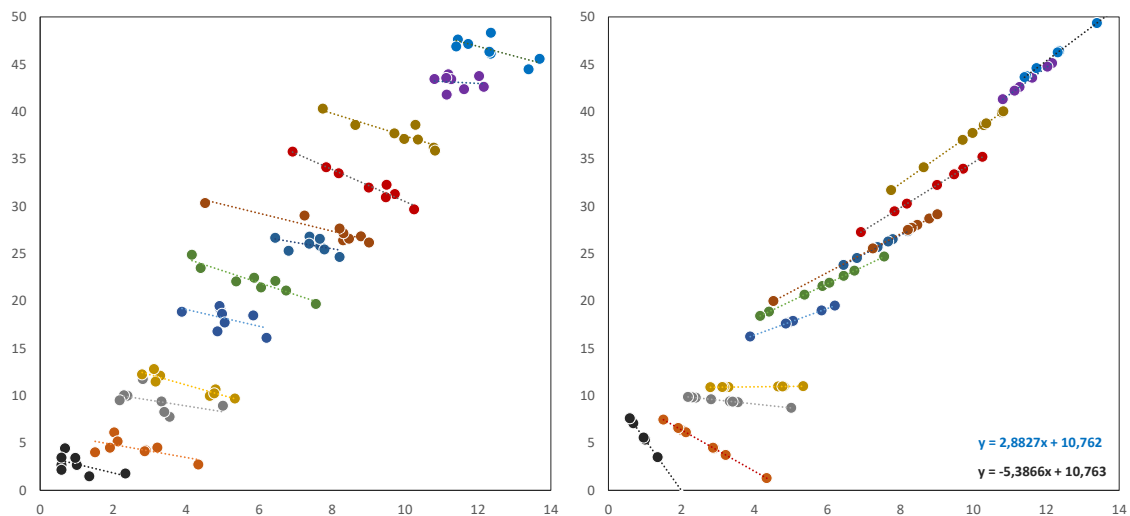
Alright, so the slope was predicted to be positive but in reality it looks like it's negative. Let's allow the model to vary the slope then, meaning each subgroup, each study, is allowed to have a different slope for the Effect.

What we are doing now is telling the model that we don't assume that everyone has the same slope. We still think the Effect exists, there is a population level effect that's the same for everyone, but each subgroup (study) shows this same effect to a different degree. In some of the studies in our meta-analysis the effect was strong, while in other studies the effect might have been weaker.

The Slope now forms a normal distribution, with the mean being the population Slope. See, we still assume there is an effect, but each study now shows a slope that falls on a distribution of Slopes.

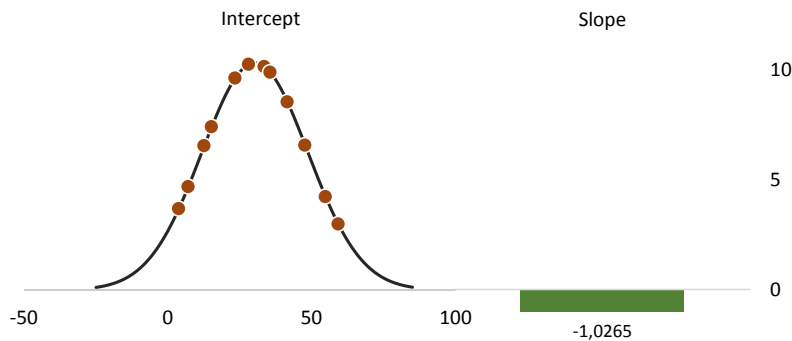


Running the model shows the prediction below, oof that doesn't look right either. There's nothing wrong with the model, it did exactly what we asked. As you can see, each sub-group has their own slope, this slope varies around the mean population slope ($M=0.848$; $SD=2.524$). The problem isn't the model, the problem is us. We made a bad assumption.



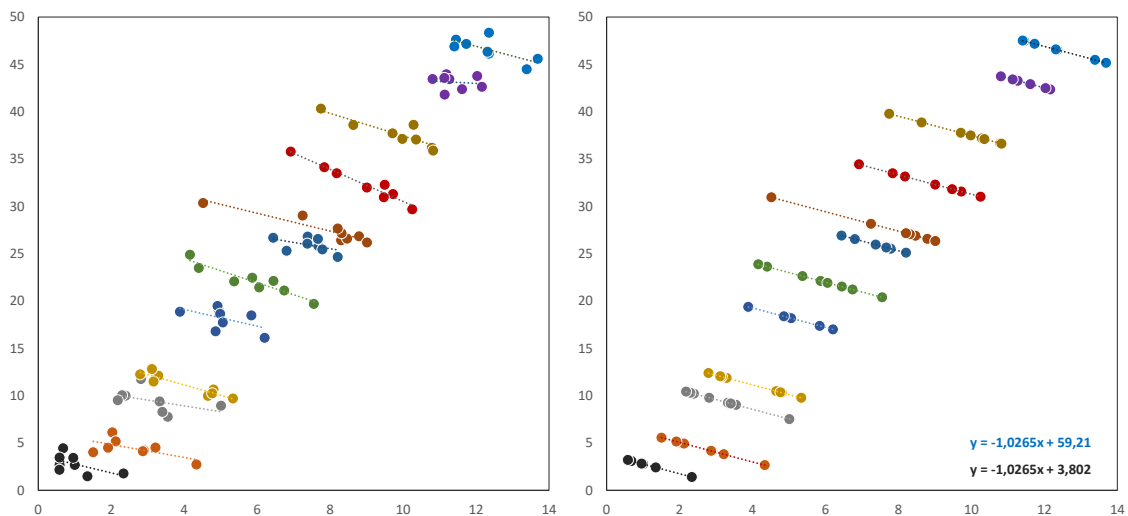
4.3.3 Model 3: Random Intercept Model

The random Slope model wasn't a huge success, what about the intercept? We can instruct the model to allow the intercept to vary. Remember, the intercept is the value of Y when X is 0. The assumption we make now is that there is a general population intercept and each Study (sub-group) has a different intercept that varies around the population intercept.



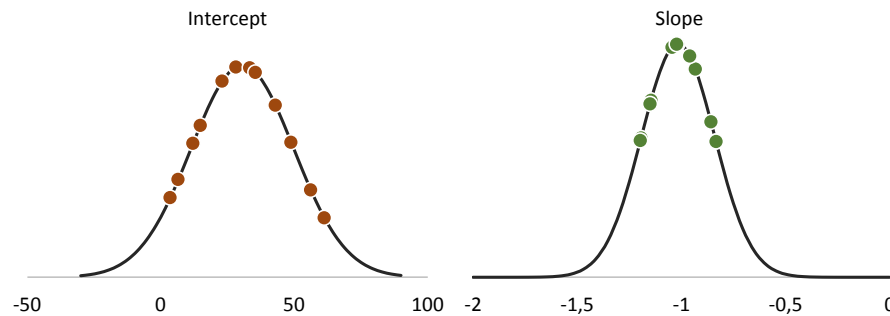
This looks better. Each Study (sub-group) has the same slope (-1.0265), but a different intercept varying around the population intercept ($M=30.226$; $SD=18.434$). Allowing the intercept to vary allows for a better estimate for the slope as well. The Effect is the same for everyone, but depending on the Study (sub-group) the start value might differ.

Often a random intercept is the first thing you should test, random slopes with fixed intercepts rarely make sense in a multi-level context, you'd need good justification for doing so.

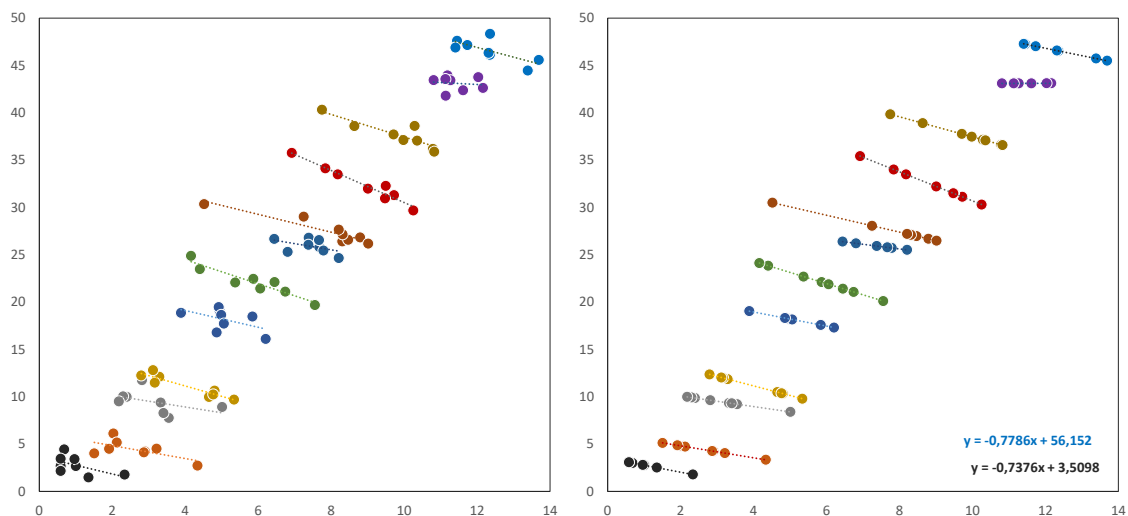


4.3.4 Model 4: Random Intercept and Slope Model

In the name of completeness, we'll also run a model where both the intercept and slope are allowed to vary between Studies (sub-groups). The intercept varies around 30.567 (SD: 19.275) and the slope varies around -1.015 (SD=0.175).



The model isn't bad, the slopes vary a little bit but as you can see the SD is pretty small so it might not be a better fit than the Random Intercept model we had before.



The model fit measures of all the models agree with me (what a surprise), the Random Intercept model has the smallest AIC and BIC value, indicating it has the best fit. By simply letting the various studies (sub-groups) vary their intercept we saw that a strong positive effect is actually composed of several negative effects. We would have drawn the exact opposite conclusion if we didn't try random effects.

	Parameters	-2LL	AIC	BIC
Fixed Model	3	568.714	570.714	573.258
Random Slope	4	545.484	549.484	554.571
Random Intercept	4	326.096	330.096	335.183
Random Intercept+Slope	6	324.455	332.455	342.628

Changing a variable from a fixed to a random effect means we assume that there's a single effect in the population (a general effect) and the levels of that variable form a sample of that same effect.

This is why it's often stated that you shouldn't turn your effect of interest into a random effect. When you are comparing for example medication A, B, and C you want to know the differences between those three. You **do not assume** A, B, and C are the same drug (*no general medication effect but a specific effect of A, B, and C*). Turning this into a random effect means you can't say anything anymore about the differences between those drugs, your assumptions changed and are no longer in line with such a conclusion

If instead you have the same drug and you test the differences between 3mg, 6mg, and 9mg you **ARE** looking at a general effect of this drug but at different levels. You don't care about the difference between 3mg and 6mg but do care about the general effectiveness of the drug at a lower or higher dosage (you could've also used 3,5mg and 6,5mg). You assume there's a single effect of this drug, your dosages are then a random sample of all possible dosages which might show the same effect more or less. In this case you already assume a random effect, so you want to conclude something about the general effectiveness of the drug. In this case turning dosage into a random effect **is** a good idea.

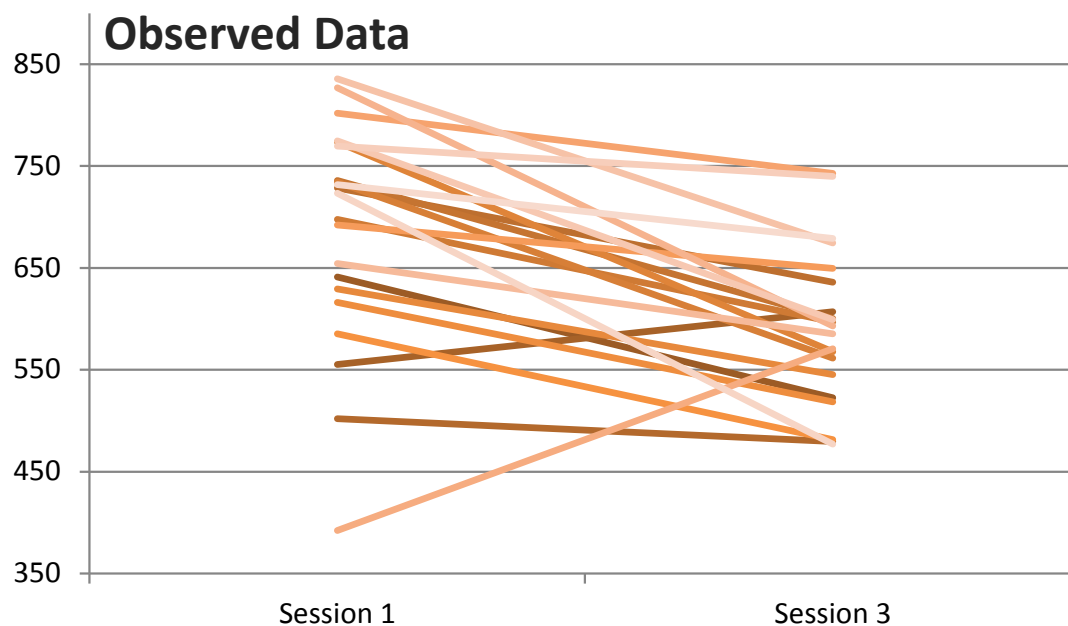
Whether or not your effect can be random depends on the assumptions you make about this effect. In most cases our effects of interest won't be suited for random effects (we often want to know differences between specific conditions), but you can see that it is certainly justifiable to have a random effect of interest (just makes interpretation a bit trickier). It's all about what you want to know, do you care about these specific conditions (Drug A, B, and C) or are you interesting in the population these conditions come from (Drug X at all dosages using a representative sample of 3mg, 6mg, and 9mg)?

4.3.5 Real Example: Random Session Effects

The example we went through just now was fairly easy and obvious, your own data will most likely not be a perfect example of a Simpson Paradox (that's where the total has one correlation but the subgroups have a correlation in the opposite direction).

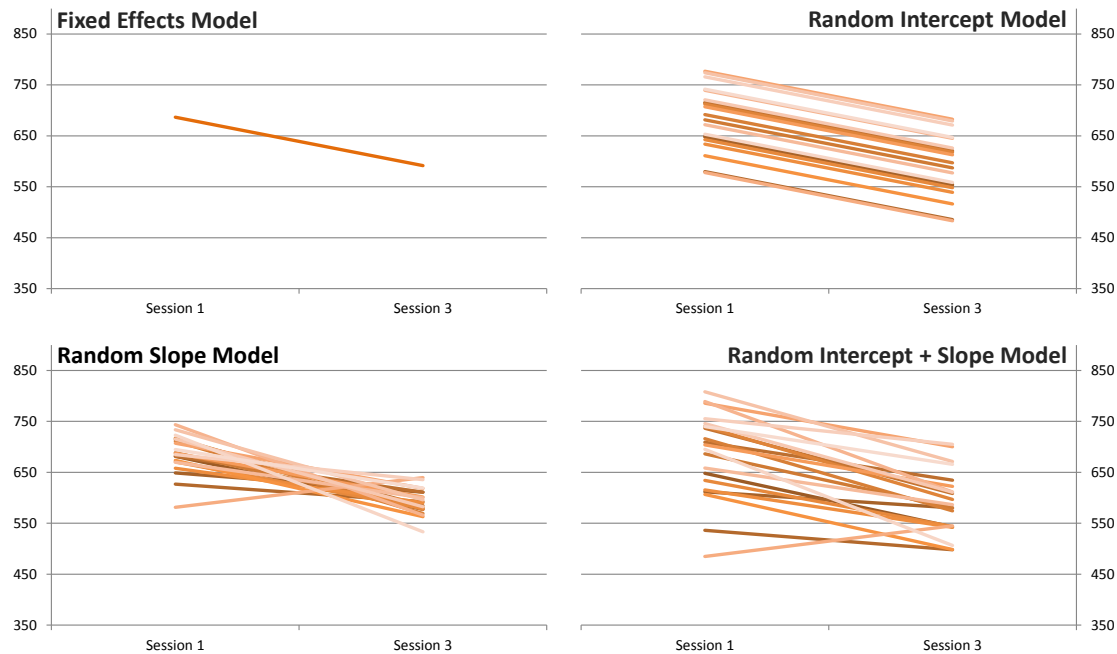
An example closer to home is the one below. This would be data that belongs to the same experiment we used before; the priming study with multiple sessions. It's behavioral data on congruent trials (dots and symbol show the same quantity). We only have the Session effect (Session 1 & Session 3), the Symbol effect (Low & High) and their interaction.

Looking at the data for the Session effect it might become obvious this data would be suitable for a random effects model. Most people have a negative slope, with a couple showing an increase in reaction time over time, they all start at a different point and end at a different point.



We can run the same four models as before, starting with a Fixed Models, and then adding a random intercept, random slope for session, and both.

- **The Fixed Effects model** doesn't seem right, ascribing the same session effect to everyone is a big assumption to make.
- **The random intercept model** is better, but not everyone is going down it doesn't fit what we are seeing.
- **The Random Slope Model** looks more like it, different people have different slopes and the estimates we'd get from this effect would more resemble the data.
- Having both a **Random Intercept and Random Slope** is an almost exact match to the observed data.




Looking at the model comparison statistics it seems that the random slope is a good addition to this model, it improves the fit by quite a bit. The additional random Intercept doesn't really do much in this case, the fits doesn't improve so it's better to stick to a random Slope model.

The significance values also change. By allowing the session effect to vary we get better estimates for each participant. This results in the interaction becoming almost significant. In the end it didn't matter that much for the conclusion, but you can imagine in more extreme cases the inclusion of random effects can alter the conclusions of your experiments.

Reporting the random session effect would mean that you have to report the mean and the standard deviation of this parameter.

Structure	Parameters	-2LL	AIC	BIC
Fixed	7	952.447	958.447	965.667
Intercept	8	949.861	957.861	967.487
Slope	8	947.084	955.084	964.711
Both	9	945.060	955.060	967.094

Type III Tests of Fixed Effects				
Source	Fixed	Intercept	Slope	Both
Intercept	.000	.000	.000	.000
Session	.000	.000	.000	.000
Symbol	.000	.000	.000	.000
Session * Symbol	.120	.176	.059	.087



5. Customizing the Analysis

Introduction

For this last session we're going to look at customization of an analysis, or more precisely a model. When running Mixed Models we have a ton of options, and the one we'll discuss here is **Custom Contrasts**. Most of our experiments aren't about overall effects, we want to know which conditions differ, if one condition is better than all the others, or if there is a trend over time.

The regular ANOVAs or regressions hook you up with pairwise comparisons, which are unplanned comparisons. These might suffice in a lot of situations, but sometimes you just want to do a comparison that isn't a pairwise comparison. Think about comparing one condition to a combination of the others or just making the one pairwise you care about. By planning your post-hoc tests you increase power and you can draw conclusions you want.

Trend analysis is also popular for repeated measures. RM ANOVA gives you these automatically but the Mixed Procedure isn't so nice. Luckily the logic of planned comparisons extends to polynomials and we can easily order trends to test them.

Finally we quickly touch upon **Generalized Estimation Equations**, which is a non-parametric regression alternative. We can use this analysis for 'ugly' data that violates assumptions.

5.1 Custom Contrasts

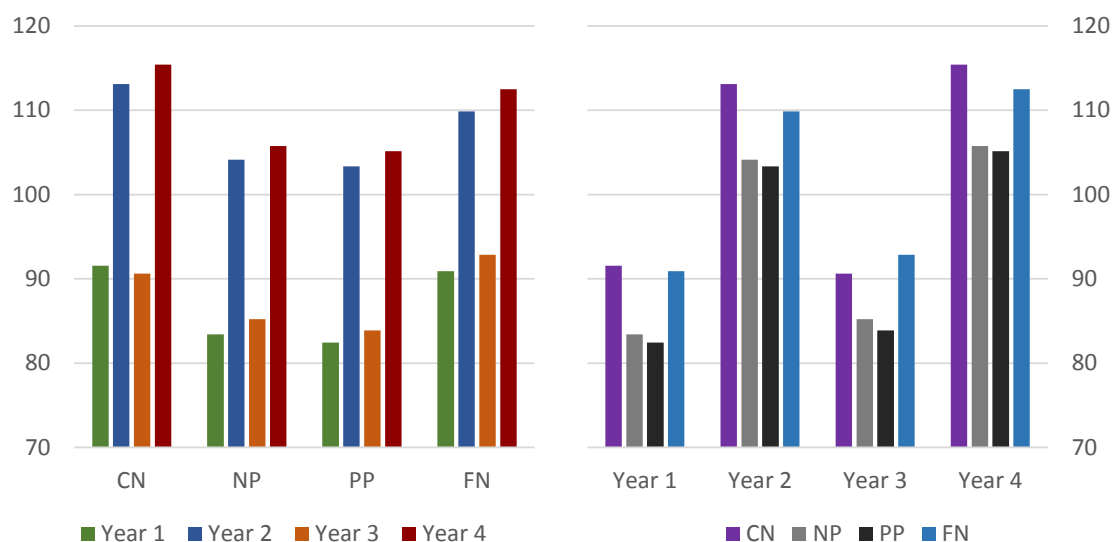
When using Repeated Measures ANOVA (GLM) or the Mixed Model Procedure, we tend to end up with a nice overall test giving us main effects, maybe an interaction or two. In reality we are more interested in the Estimated Marginal Means, we want to compare our groups to see what is causing the Main or Interaction effects.

We can proceed by ordering EMMs, we covered this in session 3, and look at the pairwise comparisons. Awesome, done, well. . . Maybe not. The first problem is the issue of multiple comparisons (Bonferroni is known, Sidak is another, but anything other than Bonferroni is good). With just a couple of groups it really doesn't matter which adjustment you use, but if you are struggling with many comparisons try out a sequential method (like Holm).

If we'd be happy with unplanned comparisons then this session would be over too fast, so we're complicating stuff. Instead of comparing our groups to all other groups, we want to **compare one group to the average of the others (L1)**, **compare the average of two groups to the average of two other groups (L2)**, and find out the **trend over Year** (polynomials). The first two scenarios are more common than you'd think, and instead of simply aggregating your data and bending over backwards to use standard methods, we simply change the current analysis to do it for us. Trend analysis for repeated measures is without a doubt something you'll want to do if you have any kind of ordinal within subject variable (like time). Customizing the analysis means you only have to run a single model instead of multiple analyses, but it also means you'll have to think a bit.

Estimates

Track	Year 1	Year 2	Year 3	Year 4
Cognitive Neuroscience	91.57	113.09	90.61	115.39
Neuropsychology	83.40	104.14	85.21	105.74
Psychopathology	82.45	103.33	83.87	105.14
Fundamental Neuroscience	90.91	109.86	92.85	112.47



To illustrate contrasts we'll be using data from four **tracks** (Between Subjects) over four **years** (Within Subjects), making it a mixed design (as in BS and WS factor, no random effects). If we look at the charts some effects might become apparent. Running the analysis shows a main effect of both Track ($p < .001$) and Year ($p < .001$), and no interaction ($p = .310$).

5.1.1 Custom Contrasts in a GLM

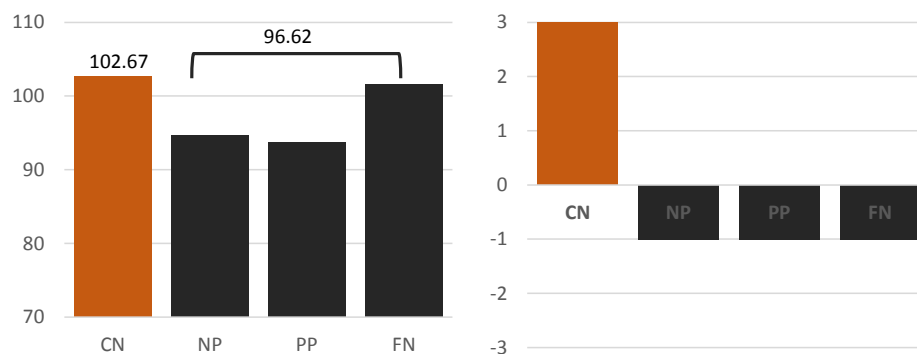
The official definition for a contrast is: A linear combination of variables (parameters or statistics) whose coefficients add up to zero.

If that doesn't make perfect sense, I don't know what will. The idea is that we have our means (the variables) and we multiply them by a contrast to get a value (the multiplication and subsequent value/sum we get is what they mean with a linear combination of variables). We end up with a weighted sum of the group means. This weighted sum of the means tells us if they conform to the pattern (the contrasts).

Before tackling the polynomials we start with the much easier to understand planned comparisons. We defined two of those:

- **Compare one group to the average of the others (L1)**
- **Compare the average of two groups to the average of two other groups (L2)**

For the first comparison we'll compare CN to the other tracks. The official contrast for such a comparison is $3/-1/-1/-1$. This means we multiply the mean of CN by 3 and all the others by -1, you add it all up and you get the difference. Mathematically this is the same as taking the average of CN and the average of (NP+PP+FN) and taking the difference (the sum of a multiplication is just faster and more efficient than summing two groups, dividing, and then taking the difference). If there is no difference between the groups this will be zero, if it's significant then it's not zero.

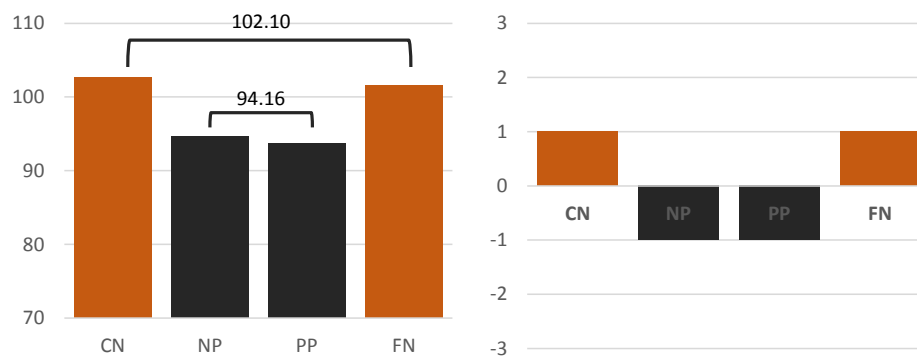


You can totally do this by hand; multiply by the appropriate coefficient and bam, done. Doing this for all participants gets you a mean and a standard error allowing for significance testing. The weighted sum for L1 is 18.16, but the difference between CN and the average of the other three is 6.05: L1 is a multiple of 3. You can undo the scaling by dividing by the largest coefficient, and voila 6.05.

The "official" contrast is multiplied to keep out fractions because not all humans math equally. It makes all the math easier to do and it doesn't change the conclusion. The mean is multiplied by three, but the standard error as well, so the significance will remain the same.

	Mean	L1		Scaled	
CN	102.67	3	308.01	1	102.67
NP	94.62	-1	-94.62	-1/3	-31.54
PP	93.70	-1	-93.70	-1/3	-31.23
FN	101.53	-1	-101.53	-1/3	-33.84
		=18.16		=6.05	

The second comparison works exactly the same, except we take the average of CN+FN and NP+PP. The coefficients here are 1/-1/-1/1, comparing the outer two with the inner two. Instead of taking the average and then the difference, it is much easier (for the computer) to multiply each group with 1/2 and -1/2 and take the sum. Like before, this can be scaled up as much as we want, the official contrast using 1 instead of 1/2.



	Mean	L2		Scaled	
CN	102.67	1	102,67	1/2	51,34
NP	94.62	-1	-94,62	-1/2	-47,31
PP	93.70	-1	-93,7	-1/2	-46,85
FN	101.53	1	101,53	1/2	50,77
		=15.88		=7.94	

Running Planned Comparisons in RM ANOVA

The theme of this course has always been comparing the RM ANOVA to the Mixed Procedure. We'll start with running planned comparisons in the regular GLM (this will work for RM ANOVA but also regression and regular ANOVAs). To run a GLM the data will need to be in a multivariate format and the analysis we'll be extending is the one below.

```
*Mixed ANOVA for Year by Track.
GLM Year1 Year2 Year3 Year4 BY Track
  /WSFACTOR=Year 4 Polynomial
  /METHOD=SSTYPE(3)
  /CRITERIA=ALPHA(.05)
  /PRINT=PARAMETER
  /WSDESIGN=Year
  /DESIGN=Track
```


Because the GLM is a more “mainstream” analysis it has multiple options pre-programmed in. The first is the `/CONTRAST` command. It allows you to choose from several coding schemes such as **DEVIATION** or **HELMERT** coding. Helmert for example compared the group to the average of the groups coming after and ignoring the ones before (reverse Helmert literally does the opposite, comparing to the ones before ignoring what comes after).

That’s nice from IBM to include, but not what we really want. The **SPECIAL** option of the `/CONTRAST` command comes closer. It allows us to specify the coefficients ourselves. These will give us the same conclusion and *p*-values, and we could be done with it... but we’re not.

```
/CONTRAST (Track) SPECIAL (3 -1 -1 -1)
/CONTRAST (Track) SPECIAL (1 -1 -1 1)
```

As nice as the `/CONTRAST` command is, it won’t accept fractions. It doesn’t matter for the result, but for completeness and control we’ll use one last method: the `/LMATRIX`. This command allows us to construct our own L-Matrix using fractions and everything.

```
/LMATRIX = "CN vs All" Track 1/2 -1/6 -1/6 -1/6
/LMATRIX = "CN/FN vs. NP/PP" Track 1/4 -1/4 -1/4 1/4
/LMATRIX = "Mean FN" Intercept 1/2 Track 0 0 0 1/2
```

That’s a long walk for a glass of water, but it’ll pan out in the end. The `/LMATRIX` command is the most flexible one we have and allows full customization, not to mention it resembles the command we need for the **Mixed Procedure** later on (and other programs). It also gives one more option you might never ever use:

```
/KMATRIX = 0
```

Without going into details, the K-matrix is what we compare the L-Matrix to. By default it is set to 0, since we expect the difference to be 0, but you can set this to anything. The output for the first contrast is shown below where we can find all the values. The **Contrast Estimate** is the weighted sum. The hypothesized value is what we set the K-Matrix to. At the bottom you can find the significance test with confidence interval. We can see that difference between CN and the average of the other groups is significant.

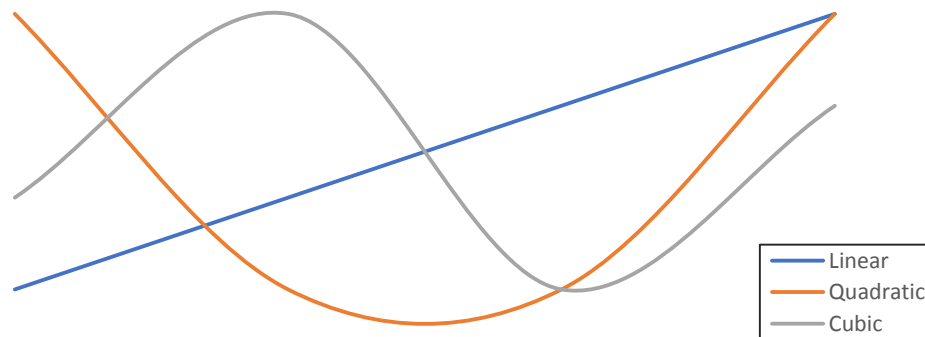
Contrast Results (K Matrix) ^a		
Contrast		Averaged Variable
		MEASURE_1
L1	Contrast Estimate	6,052
	Hypothesized Value	,000
	Difference (Estimate - Hypothesized)	6,052
	Std. Error	,513
	Sig.	,000
	95% Lower Bound	5,034
	Confidence Upper Bound	
	Interval for Difference	7,070

a. Based on the user-specified contrast coefficients (L') matrix: CN vs All

Running Polynomial Contrasts in RM ANOVA

A Polynomial contrast is a special version of contrasts and a form of trend analysis. It more or less works like the contrasts before. We put in the contrast and we expect this to be zero, but in this case zero means that the trend doesn't exist. You can only do this with ordinal data that has equally spaced levels (like income, education, or year).

The way polynomials work is a whole math thing, which doesn't matter right now. What is important is that you can test polynomials up to an order that is one less than the number of levels you have. A linear trend for example is a first order polynomial (a straight line), and to make a line you need at least two points. The second important thing is that the highest order polynomial wins. If there is a significant quadratic trend then the significant linear trend is no longer valid (it can't be a straight line if an inverted-U describes it better).



The regular GLM has polynomials baked in since it's a very common test for repeated measures. It gives you a nice new table called Test of Within-Subjects Contrasts.

```
/WSFACTOR=Year 4 Polynomial
```

Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	Year	Type III Sum of Squares	df	Mean Square	F	Sig.
Year	Linear	11680,367	1	11680,367	430,064	,000
	Quadratic	26,619	1	26,619	1,116	,293
	Cubic	32816,392	1	32816,392	1134,184	,000
Year * Track	Linear	1,232	3	,411	,015	,997
	Quadratic	44,033	3	14,678	,615	,607
	Cubic	224,403	3	74,801	2,585	,058
Error(Year)	Linear	2607,322	96	27,160		
	Quadratic	2289,828	96	23,852		
	Cubic	2777,657	96	28,934		

The table tells us there's a significant linear and cubic effect for Year and an almost cubic trend for the interaction (which tests the trend per level Track). Because the cubic trend is significant the linear trend is no longer valid.

5.1.2 Custom Contrasts in the Mixed Procedure

The whole idea of the course is to do this using the Mixed Procedure though, which is a bit trickier. The Mixed Procedure doesn't come with those nice baked in functions like the GLM, we're going to have to do it ourselves using the `/TEST` command. We'll have to specify all the coefficients, including those of the interactions, Mixed isn't helping you at all.

The `/TEST` command works the same as the `/LMATRIX` command we used for the GLM (see there was a reason for using those). We can set the hypothesized value and even use fractions, but we don't have to. Omitting the hypothesized value will default to zero and instead of using fractions we can add the `DIVISOR` parameter which divides all coefficients by that value (makes it easier for us humans).

```
/TEST(Hypothesis)="name of the contrast" [Variable] 3 -1 -1 -1 DIVISOR=3
```

Main Effects Only

We'll start easy, with a model that only contains the main effects. We can order the same CN vs All contrast as before with the `/TEST` command.

```
*Custom Contrast for the Main Effect of Track.
MIXED Score BY Track Year
  /CRITERIA=CIN(95) MXITER(100) MXSTEP(10) SCORING(1)
    SINGULAR(0.000000000001) HCONVERGE(0, ABSOLUTE)
    LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
  /FIXED= Track Year| SSTYPE(3)
  /METHOD=REML
  /REPEATED=Year| SUBJECT(PP) COVTYPE(UN)
  /PRINT=LMATRIX SOLUTION
  /TEST="Contrast CN vs All" Track 3 -1 -1 -1 DIVISOR=3
  /TEST="Contrast CN/FN vs NP/PP" Track 1 -1 -1 1 DIVISOR=2.
```

The output will look like the table below. We can see that there is a significant difference between CN and the rest, but the Estimate is a little bit off... that's because we didn't include the interaction while the GLM did.

Contrast Estimates^{a,b}

Contrast	Estimate	Std. Error	df	Test Value	t	Sig.
L1	6,244923	,504375	96,000	0	12,382	,000

With Interaction Effects

Now to add the interaction effects back into the model, ordering the same contrast with the `/TEST` command. We want to check CN against all other groups, this means we need to average them over each level of Year. The average of four years is the same as multiplying each by $1/4$ and taking the sum, which is exactly what we'll be doing.

I find it easier to make a table that follows the interaction. Because we have $\text{Track} \times \text{Year}$ we put track as the rows and years as the columns. I do this because the coefficient Matrix we're going to build follows the interaction, changing Year and then Track (it goes $[\text{Track}=\text{CN}; \text{Year}=3]$ - $[\text{Track}=\text{CN}; \text{Year}=4]$ - $[\text{Track}=\text{NP}; \text{Year}=1]$). By making the table like this we can simply copy the inner cells.

For the **Track** contrast I'll use the CN vs All comparison, which had coefficients $3/-1/-1/-1$. It's a main effect so we average over all the years. We want each year to contribute $1/4$ th to the estimate and we want CN to be 3 and the rest -1. CN in year one is therefore $1/4$ th of 3 (which is 0.75 or $3/4$ th) and the other tracks all have $1/4$ th of -1 (which is -0.25 or $-1/4$ th).

The coefficients for the **interaction** are the **inner cells**, which we can calculate by multiplying the left and top rows ($3 \times 1/4 = 3/4$) in the table below. The sum of the rows gives us the coefficients for the **Track main effect** and the sum of the columns gives the coefficients for the **Year main effect**. For this to work it all has to add up to zero, but that shouldn't be an issue if you put some thought in it.

		Year 1	Year 2	Year 3	Year 4	
		$1/4$	$1/4$	$1/4$	$1/4$	
CN	3	$3/4$	$3/4$	$3/4$	$3/4$	3
NP	-1	$-1/4$	$-1/4$	$-1/4$	$-1/4$	-1
PP	-1	$-1/4$	$-1/4$	$-1/4$	$-1/4$	-1
FN	-1	$-1/4$	$-1/4$	$-1/4$	$-1/4$	-1
		0	0	0	0	

You could copy these into the `/TEST` command right away, it doesn't care about fractions, but it might be more difficult for humans to understand. We can multiply everything by 4 to get rid of the fractions and use the **DIVISOR** parameter to divide it all during the analysis. I used **DIVISOR=12** because that sets the largest group coefficient to 1.

		Year 1	Year 2	Year 3	Year 4	
		$1/4$	$1/4$	$1/4$	$1/4$	
CN	12	3	3	3	3	12
NP	-4	-1	-1	-1	-1	-4
PP	-4	-1	-1	-1	-1	-4
FN	-4	-1	-1	-1	-1	-4
		0	0	0	0	

The `/TEST` command will take these values and constructs an L-Matrix for the statistical test. It doesn't really care about how it looks so you can put it all on a single line, but I can advise some structure (especially when you get more factors or three-way interactions).

Another recommendation is to write down every factor in your model (all main effects and interaction effects), at least for the first time. I did the same below, adding in the Year variable with all zeroes. Coefficients are zero by default so we can remove the year row and it will still work, but it's easy to miss one in larger models (plus it makes your syntax look really impressive).

```
*Custom Contrast for the Full Factorial Model.
MIXED Score BY Track Year
  /CRITERIA=CIN(95) MXITER(100) MXSTEP(10) SCORING(1)
    SINGULAR(0.000000000001) HCONVERGE(0, ABSOLUTE)
    LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
  /FIXED= Track Year Track*Year| SSTYPE(3)
  /METHOD=REML
  /REPEATED=Year| SUBJECT(PP) COVTYPE(UN)
  /PRINT=LMATRIX SOLUTION
  /TEST="Contrast CN vs all" Track 12 -4 -4 -4
                                Year 0 0 0 0
                                Track*Year 3 3 3 3
                                           -1 -1 -1 -1
                                           -1 -1 -1 -1
                                           -1 -1 -1 -1 DIVISOR=12.
```

The output looks more like the GLM now, in fact it is exactly the same as the GLM output and if we would manually calculate the difference. It might seem cumbersome but this command really shines when you use it to its full potential. We can construct any comparison, even within the interaction. Pairwise comparisons allow you to compare levels within the levels of another (compare groups per year of years per groups), custom contrasts allow you to compare levels of one between levels of the other (even combinations of Track in Year 1 to combinations of Track in Year 2).

Contrast Estimates^{a,b}

Contrast	Estimate	Std. Error	df	Test Value	t	Sig.
L1	6,051635	,512898	96,000	0	11,799	,000

The construction of contrasts can still be tricky though, it all has to add up and with large models you might lose sight of what the program is doing. If you get an error you can always order the L-Matrix using `/PRINT=LMATRIX` to see which coefficients are given to which effects (sometimes you switch rows and columns without realizing). Even if the contrast doesn't run the L-Matrix should still show up. The matrix on the left didn't run and I can see why. The coefficients for the Tracks are right but those for the interaction aren't, it compares year 1 to the rest for each track (I switched rows and columns). The matrix on the right did run. Printing the matrix can help you spot mistakes and see what the order of coefficients is.

Contrast Coefficients ^{a,b}		L1
Fixed Effects	Intercept	0
	[Track=1]	1
	[Track=2]	-,333
	[Track=3]	-,333
	[Track=4]	-,333
	[Year=1]	0
	[Year=2]	0
	[Year=3]	0
	[Year=4]	0
	[Year=1] * [Track=1]	,250
	[Year=2] * [Track=1]	-,083
	[Year=3] * [Track=1]	-,083
	[Year=4] * [Track=1]	-,083
	[Year=1] * [Track=2]	,250
	[Year=2] * [Track=2]	-,083
	[Year=3] * [Track=2]	-,083
	[Year=4] * [Track=2]	-,083
	[Year=1] * [Track=3]	,250
	[Year=2] * [Track=3]	-,083
	[Year=3] * [Track=3]	-,083
	[Year=4] * [Track=3]	-,083
	[Year=1] * [Track=4]	,250
	[Year=2] * [Track=4]	-,083
	[Year=3] * [Track=4]	-,083
	[Year=4] * [Track=4]	-,083

a. Contrast CN vs all

b. This L matrix is not estimable, test results will not be produced.

Contrast Coefficients ^a		L1
Fixed Effects	Intercept	0
	[Track=1]	1
	[Track=2]	-,333
	[Track=3]	-,333
	[Track=4]	-,333
	[Year=1]	0
	[Year=2]	0
	[Year=3]	0
	[Year=4]	0
	[Year=1] * [Track=1]	,250
	[Year=2] * [Track=1]	,250
	[Year=3] * [Track=1]	,250
	[Year=4] * [Track=1]	,250
	[Year=1] * [Track=2]	-,083
	[Year=2] * [Track=2]	-,083
	[Year=3] * [Track=2]	-,083
	[Year=4] * [Track=2]	-,083
	[Year=1] * [Track=3]	-,083
	[Year=2] * [Track=3]	-,083
	[Year=3] * [Track=3]	-,083
	[Year=4] * [Track=3]	-,083
	[Year=1] * [Track=4]	-,083
	[Year=2] * [Track=4]	-,083
	[Year=3] * [Track=4]	-,083
	[Year=4] * [Track=4]	-,083

a. Contrast CN vs all

Polynomial Contrast in the Mixed Procedure

With custom contrasts out of the way we can proceed to Polynomial contrasts. Much like the GLM there isn't a big difference compared to custom contrast in how they work. Instead of coding a **difference** we can code a **trend**.

Main effect Polynomials

The trend of the main effect for year is the same as we did for track before, except now we fill in the linear polynomial for Year and take the average over Tracks. I applied the divisor of 12 to keep the scale the same, but that's optional (since the significance test is not affected). We then do the same for the quadratic and cubic effects (because we have four levels) and done. I omitted the zeroes for Track to keep the code short.

		Year 1	Year 2	Year 3	Year 4	
		-12	-4	4	12	
CN	1/4	-3	-1	1	3	0
NP	1/4	-3	-1	1	3	0
PP	1/4	-3	-1	1	3	0
FN	1/4	-3	-1	1	3	0
		12	-4	4	12	

```
*Polynomial Contrasts for the Year variable.
MIXED Score BY Track Year
  /CRITERIA=CIN(95) MXITER(100) MXSTEP(10) SCORING(1)
    SINGULAR(0.000000000001) HCONVERGE(0, ABSOLUTE)
    LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
  /FIXED= Track Year Track*Year| SSTYPE(3)
  /METHOD=REML
  /REPEATED=Year| SUBJECT(PP) COVTYPE(UN)
  /PRINT=LMATRIX SOLUTION
  /TEST="Linear Year" Year -12 -4 4 12
    Track*Year -3 -1 1 3
    -3 -1 1 3
    -3 -1 1 3
    -3 -1 1 3 DIVISOR=12.
  /TEST="Quadratic Year" Year 4 -4 -4 4
    Track*Year 1 -1 -1 1
    1 -1 -1 1
    1 -1 -1 1
    1 -1 -1 1 DIVISOR=4.
  /TEST="Cubic Year" Year -4 12 -12 4
    Track*Year -1 3 -3 1
    -1 3 -3 1
    -1 3 -3 1
    -1 3 -3 1 DIVISOR=12.
```

Each polynomial is its own contrast so we'll get separate tables, if you named them it should be easy to identify them. Unlike the GLM these give us t-test between hypothesis (0) and the estimate (the GLM did F-tests), but the significance is still the same. We can determine that we have a significant cubic effect of time.

Contrast Estimates

Contrast	Estimate	Std. Error	df	Test Value	t	Sig.
L1	16,110981	,776883	96,000	0	20,738	,000

Linear Year

Contrast Estimates

Contrast	Estimate	Std. Error	df	Test Value	t	Sig.
L1	1,031864	,976778	96,000	0	1,056	,293

Quadratic Year

Contrast Estimates

Contrast	Estimate	Std. Error	df	Test Value	t	Sig.
L1	27,004688	,801858	96,000	0	33,678	,000

Cubic Year

Interaction Polynomials

We can crank things up to eleven and do one more thing. If you run a GLM with multiple within subject variables you will also see polynomials for those interactions. If we pretend Track is also a Within-Subject variable the GLM will give us output like the table below (I removed the error terms to keep it small). It contains the Track*Year interaction, where both can have a different trend and it basically tests which combination of trends is significant.

Tests of Within-Subjects Contrasts

Source	Track	Year	Type III Sum of Squares	df	Mean Square	F	Sig.
Track	Linear		94,413	1	94,413	2,943	,099
	Quadratic		6299,437	1	6299,437	347,178	,000
	Cubic		13,296	1	13,296	,964	,336
Year		Linear	11680,367	1	11680,367	497,899	,000
		Quadratic	26,619	1	26,619	,932	,344
		Cubic	32816,392	1	32816,392	1164,347	,000
Track * Year	Linear	Linear	,727	1	,727	,020	,889
		Quadratic	15,913	1	15,913	,555	,463
		Cubic	182,472	1	182,472	5,960	,022
	Quadratic	Linear	,001	1	,001	,000	,996
		Quadratic	21,960	1	21,960	1,111	,302
		Cubic	4,285	1	4,285	,166	,688
	Cubic	Linear	,503	1	,503	,019	,890
		Quadratic	6,160	1	6,160	,334	,569
		Cubic	37,646	1	37,646	1,213	,282

Running an interaction polynomial is just as easy (or difficult) as running any other polynomial. We put in the coefficients for all factors involved and run the model. I'll show you with the Linear Track*Cubic Year interaction. The contrast results in zeroes for the main effects so I omitted those. Running the model nets the same result as the GLM above, no surprise there.

		Year 1	Year 2	Year 3	Year 4	
		-1	3	-3	1	
CN	-3	3	-9	9	-3	0
NP	-1	1	-3	3	-1	0
PP	1	-1	3	-3	1	0
FN	3	-3	9	-9	3	0
		0	0	0	0	

```

/TEST="Linear*Cubic Interaction" Track*Year 3 -9 9 -3
                                         1 -3 3 -1
                                         -1 3 -3 1
                                         -3 9 -9 3 DIVISOR=9

```

Contrast Estimates^{a,b}

Contrast	Estimate	Std. Error	df	Test Value	t	Sig.
L1	-6,003653	2,459156	24,000	0	-2,441	,022

Linear Track * Cubic Year Interaction

To make this example work I did change the data, instead of 100 people going through four year in four different groups it had to be 25 people going through four years four times in different groups. The multivariate data structure of the GLM makes it impossible to accidentally run a Between Subject variable as a Within subject Variable, while the univariate data structure of the Mixed Procedure doesn't really care. The Mixed Procedure will run and give you the output below. The estimate is the same but the DF is a lot bigger because the Person effect is from 100 people and not 25.

It's a detail you have to keep in mind, you can run trends on between subject variables, but it just doesn't make any theoretical sense to do so. The GLM protects against this mistake by forcing a multivariate datafile and blocking the option of doing a Polynomial on BS-Factors.

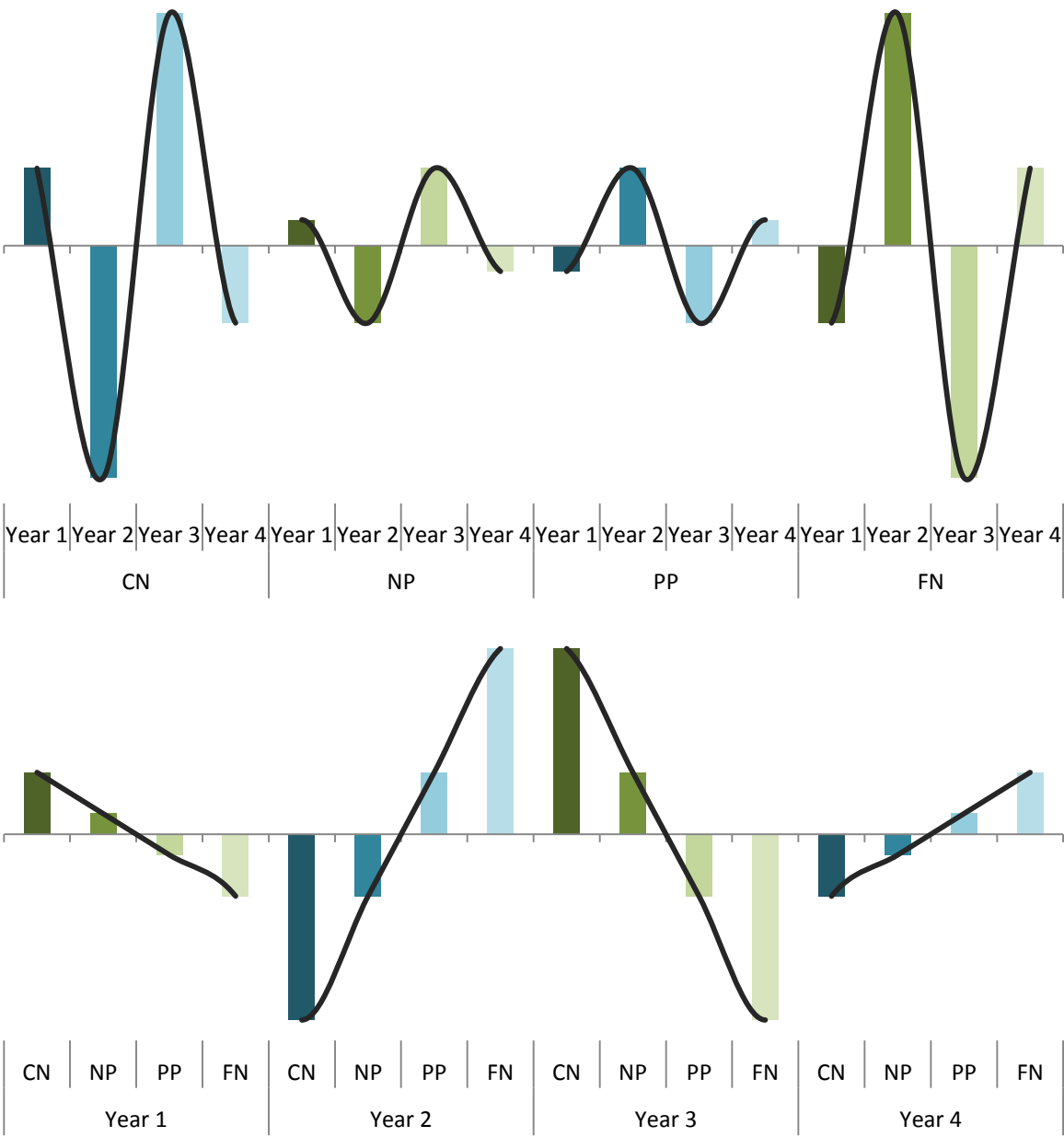
Contrast Estimates^{a,b}

Contrast	Estimate	Std. Error	df	Test Value	t	Sig.
L1	-6,003653	2,390678	96,000	0	-2,511	,014

Linear Track * Cubic Year Interaction

We've reproduced the WS*WS interaction of the GLM and it'll definitely make your analyses look a lot cooler, but you might be wondering what a Linear*Cubic interaction even is. A picture says more than a thousand words so I put the Linear*Cubic interaction on the next page. In both graphs Year is a cubic effect and Track a Linear effect.

In the top graph the Cubic effect of Year changes from large, to small, to small reversed, to large reversed (a linear change in the cubic year effect). In the bottom graph the linear effect of Track changes from small negative, to large positive, to large negative, to small positive (a cubic change in the linear Track effect).



Multiple comparisons

When using contrasts, you will need to adjust for multiple testing manually. Each contrast can be seen as a pairwise comparison, if you have three contrasts you would need to adjust for three pairs. Two of these can be calculated by hand: the Bonferroni and Sidak adjustment. In practice you will see little difference between the two on low number of pairs though. You can either adjust the alpha or the p-value (not both!).

$$\begin{array}{l|l} \alpha_{Bonferroni} = \alpha/k & p_{Bonferroni} = p * k \\ \alpha_{Sidak} = 1 - (1 - \alpha)^{1/k} & p_{Sidak} = (1 - (1 - p)^k) \end{array}$$

In these equations α stands for the alpha used, which is usually 0.05 (95%) and K stands for the number of pairs you have. In the example of three contrasts and the p-value of .009 we would get an alpha of .017 or a p-value of .027, either way it is still significant.

$$\begin{array}{l|l} \alpha_{Bonferroni} = 0.05/3 = .017 & p_{Bonferroni} = .009 * 3 = .027 \\ \alpha_{Sidak} = 1 - (1 - 0.05)^{1/3} & p_{Sidak} = (1 - (1 - .009)^3) \end{array}$$

For comparisons adjustment for multiple testing is needed for sure, when it comes to polynomials it's a bit fuzzier. Most don't seem to be doing this, SPSS for example shows unadjusted p-values (but I'm sure someone will disagree).

For a different adjustment I tend to copy unadjusted p -values and run them through an R-script to get adjusted p -values. Which method you use depends on your data, the reason why SPSS limits the options to **Bonferroni** and **Sidak** is because those two make less assumptions than the others. Between them, **Sidak** assumes independence of the tests while **Bonferroni** doesn't.

For *Unplanned Comparisons* the **Tukey** method might be more appropriate, since it was developed to adjust when comparing every mean to every other mean (e.g. pairwise comparisons). If you have a reference group and compare all means to that group the **Dunnett** method was made to adjust for just that. There are many more and depending on your specific set-up a different adjustment might be more appropriate. There are also sequential (**Holm**) methods that use a step-down approach, testing the biggest effect first and adjusting until there's no more significance.

Bonferroni is a catch-all; it assumes no independence and is the one every knows. Picking another method requires some explanation that reviewers or supervisors might not accept (especially sequential methods, those just sound like cheating don't they?).

5.2 Generalized Estimation Equations (GEE)

The final note of this course introduces a new type of analysis, the **Generalized Estimation Equation** (GEE). The GEE is a Generalized Linear Model and it can accommodate auto-correlated data (dependent data such as repeated measures) and non-normal data. This is a fancy way of saying that GEEs can deal with ugly data. By ugly data I mean data that isn't normally distributed and cannot be transformed in any way to be normal. Ugly data is a problem because all of the parametric tests assume normality, they use means and SDs which doesn't work if the data is skewed.

I often run into ugly data when analyzing accuracy scores. Unless your experiment has been set-up for it, accuracy will probably be heavily skewed to the left (most people will have a high accuracy and the tail is to the left). We do this on purpose of course, we often aim for 80% accuracy to have enough valid trials and some invalid trials to compare to.

A less common example in our line of work is count-data, which means the data can only take on non-negative integers. It's literally when you count things, there can never be less than 0 cells for example and there can also never be two and a half cells. Count-data tends to follow a Poisson distribution, which means that the odds of a given count depends on the interval of counting. Regular tests are probably not appropriate for this data, the assumptions won't hold.

There's a way to analyze anything, and here we'll discuss the **GEE**. The nitty gritty of GEEs is way to complicated, I don't know all the math and details either (I just look up and understand the parts I need to use them).

5.2.1 Running a GEE in SPSS

The GUI of the GEE look a bit different than we're used to and that's mostly because IBM is being fancy. The way we run a GEE is more or less the same as the Mixed Procedure, with a few extra options.

Repeated: in the repeated tab we can select the subject variables and repeated variables. Below that you can find the "Structure" where we can specify the correlation matrix (variance-covariance matrix). This is the same as the first window of the Mixed Procedure.

Response: here we put in the dependent variable, not much to it really.

Predictors: this is where you add independent variables as either factors (categorical) or covariates (continuous). The Options allow you to choose the reference category (first or last).

Model: specify your model here, add the effects by adding in main effects and interactions (or select factorial and add everything all at once).

EM Means: exactly what you expect, here we can specify for which terms we want estimated marginal means. We can also specify the contrast here, popular are Pairwise and polynomials. You're not limited to one EMM per term, you can add terms multiple times and request a different contrast each time. *Note: the pairwise for interactions will by default compare each combination of the terms with each other combination, if you want to do pairwise comparisons per level of the other variable, you'll have to specify that in the syntax.*

For the **Track*Year** analysis we did in the last section on Contrast we'd get something like this. It's regular data with a normal distribution, so we don't need to take any ugliness into account. We run main effect and the interaction plus order EMMs. Unfortunately, the base-version of GEE doesn't allow for custom contrasts (that requires an expansion module IBM is happy to sell you), but other programs like R will be more flexible (the same logic applies).

```
*GEE analysis for the Track*Year Data.
GENLIN Score BY Track Year (ORDER=ASCENDING)
/MODEL Track Year Track*Year INTERCEPT=YES DISTRIBUTION=NORMAL
      LINK=IDENTITY
/CRITERIA SCALE=MLE PCONVERGE=1E-006 (ABSOLUTE) SINGULAR=1E-012
      ANALYSISTYPE=3 (WALD) CILEVEL=95 LIKELIHOOD=FULL
/EMMEANS TABLES=Track SCALE=ORIGINAL COMPARE=Track CONTRAST=PAIRWISE
      PADJUST=LSD
/EMMEANS TABLES=Track SCALE=ORIGINAL COMPARE=Track CONTRAST=HELMERT
      PADJUST=LSD
/EMMEANS TABLES=Year SCALE=ORIGINAL COMPARE=Year CONTRAST=POLYNOMIAL
      PADJUST=LSD
/EMMEANS TABLES=Track*Year SCALE=ORIGINAL COMPARE=Track*Year
      CONTRAST=PAIRWISE PADJUST=LSD
/EMMEANS TABLES=Track*Year SCALE=ORIGINAL COMPARE=Track
      CONTRAST=PAIRWISE PADJUST=LSD
/EMMEANS TABLES=Track*Year SCALE=ORIGINAL COMPARE=Year
      CONTRAST=PAIRWISE PADJUST=LSD
/REPEATED SUBJECT=PP WITHINSUBJECT=Year SORT=YES CORRTYPE=UNSTRUCTURED
      ADJUSTCORR=YES COVB=ROBUST
/MISSING CLASSMISSING=EXCLUDE
/PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

The output is a bit different though, we're no longer running a parametric analysis so the regular statistics such as F-values won't do. Like the Mixed Procedure it starts with model information, a summary of the model. After the model Information table you'll find a couple more summary tables showing you the number of levels, descriptives, and valid cases.

Model Information

Dependent Variable	Score
Probability Distribution	Normal
Link Function	Identity
Subject Effect 1	PP
Within-Subject Effect 1	Year
Working Correlation Matrix Structure	Unstructured

The next important table is the **Goodness of Fit table**, fulfilling the same role as the same table in the Mixed Procedure. The difference here is that -Log Likelihood (and the derived AIC) doesn't work for this model, instead there's a different measure called QIC that is valid and works just like an AIC value. Smaller is better and comparing QIC between different models allows you to specify the correct parameters. QIC is used to test between correlation structures using the same model terms, QICC is used to decide between models with different terms using the same correlation structure. Nothing new here, it's just like Mixed but with different measures, except we don't need to switch between ML and REML.

Goodness of Fit		Tests of Model Effects			
	Value	Type III			
		Wald Chi-Square	df	Sig.	
Quasi Likelihood under Independence Model Criterion (QIC)	9600,873				
Corrected Quasi Likelihood under Independence Model Criterion (QICC)	9600,873				
Source					
(Intercept)		203347,974	1	,000	
Track		332,593	3	,000	
Year		1973,460	3	,000	
Track * Year		11,571	9	,239	

Then we get to the main part, the **Tests of Model Effects**. This is your overall test for predictor significance. It doesn't use F-tests since it's non-parametric, instead it used a Chi-Square test. We can see that we have Main effects and no interaction, perfect. The **Parameter Estimates** table shows the coefficients that are used to calculate the fitted values. These also have significance values, these help with interpretation, but shouldn't be reported unless you have covariates (continuous variables). For a proper conclusion we look at estimated marginal means that take all the effects into account.

Parameter Estimates							
Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	112,475	,7633	110,979	113,971	21712,126	1	,000
[Track=1]	2,916	1,2110	,542	5,290	5,798	1	,016
[Track=2]	-6,739	1,2501	-9,189	-4,289	29,064	1	,000
[Track=3]	-7,334	1,0897	-9,470	-5,198	45,294	1	,000
[Track=4]	0
[Year=1]	-21,561	1,2210	-23,954	-19,168	311,801	1	,000
[Year=2]	-2,611	1,5967	-5,741	,518	2,674	1	,102
[Year=3]	-19,627	1,3446	-22,262	-16,991	213,068	1	,000
[Year=4]	0
[Track=1] * [Year=1]	-2,259	1,9883	-6,156	1,638	1,290	1	,256
[Track=1] * [Year=2]	,313	2,2136	-4,026	4,651	,020	1	,888
[Track=1] * [Year=3]	-5,156	1,9680	-9,013	-1,299	6,864	1	,009
[Track=1] * [Year=4]	0
[Track=2] * [Year=1]	-,776	1,9023	-4,505	2,952	,167	1	,683
[Track=2] * [Year=2]	1,011	2,0973	-3,099	5,122	,233	1	,630
[Track=2] * [Year=3]	-,897	1,8445	-4,512	2,719	,236	1	,627
[Track=2] * [Year=4]	0
[Track=3] * [Year=1]	-1,126	1,6985	-4,455	2,204	,439	1	,508
[Track=3] * [Year=2]	,796	2,1170	-3,353	4,945	,141	1	,707
[Track=3] * [Year=3]	-1,649	1,7742	-5,126	1,829	,864	1	,353
[Track=3] * [Year=4]	0
[Track=4] * [Year=1]	0
[Track=4] * [Year=2]	0
[Track=4] * [Year=3]	0
[Track=4] * [Year=4]	0
(Scale)	24,919						

Among the various EMMs I also ordered one for Year with **Polynomial contrasts**. Just like the **GLM** and the **Mixed Procedure** it says we have a *Cubic effect*, isn't that amazing?

Individual Test Results					
Year Polynomial Contrast	Contrast Estimate	Std. Error	Wald Chi-Square	df	Sig.
Linear	10,8076	,51062	447,983	1	,000
Quadratic	,5159	,47852	1,162	1	,281
Cubic	18,1153	,52703	1181,442	1	,000

5.2.2 Link Functions

In the **Type of Model** tab we can specify the distribution and the link function, which is at least one new word. A link function is like a transformation, which is applied to the data in order to make it fit in a linear model. There are a few options to choose from here, conveniently grouped together in categories based on the type of response. The options are composed of a distribution and a link function, we have a couple of options for these two.

Distributions


- **Binomial**: this is for binary data, yes/no or present/absent.
- **Gamma or Inverse Gaussian**: this is appropriate for positive data skewed to larger positive values (like accuracy data).
- **Negative Binomial**: to make a long story short, a binomial distribution is "the number of trials required to get an x-amount of successes". It's for positive integer values.
- **Normal**: this is the familiar one and assumes the good old bell-shaped numeric distribution with a central mean.
- **Poisson**: that's the one I mentioned before, it's "number of occurrences in a fixed period of time". This would for example work with cell-cultures and you're counting the number of cells, only positive integers.
- **Tweedie**: this is a mix between continuous and discrete distributions, the data has to be 0 or larger.
- **Multinomial**: the last one is for ordinal data, think likert-scale values (agreement or even SES).

Link Functions

There are a couple of these, won't go into them all. We really only need to look at three of them, all the others only work for binomial or multinomial distributions and we're not going into that.

- **Binomial**: this is for binary data, yes/no or present/absent.
- **Identity**: $Y=X$
- **Log**: $Y=\text{Log}(X)$
- **Power**: $Y=X^a$ where a is not 0 (if it's zero you get the log)

Most of what we discussed for **Mixed Models** applies to **GEEs** as well, except a GEE can handle different data. You can test count-data, non-normal accuracy data, even dichotomous 0/1 data. Unfortunately it doesn't support Random effects. Ugly data with random effects can still be analyzed using the **Generalized Linear Model**, but doing that in SPSS is a bit weird because of the "fancy" interface. The logic still applies and you should be able to run these models with what you know now.



6. Appendices

Introduction

I added some appendices to the end, it contains a coefficient table for polynomials up to 10 levels, though generally you won't need more than 4 (more levels might be better as a continuous variable). Appendix B contains a very short overview of the Mixed procedure, the assumptions, options in case of violation, how to run it in SPSS (menus and syntax), and the steps you need to take.

6.1 Appendix A: Polynomial Contrast Coefficients

Polynomial Contrasts Coefficients

2 Levels										
Linear	-1	1								
3 Levels										
Linear	-1	0	1							
Quadratic	1	-2	1							
4 Levels										
Linear	-3	-1	1	3						
Quadratic	1	-1	-1	1						
Cubic	-1	3	-3	1						
5 Levels										
Linear	-2	-1	0	1	2					
Quadratic	2	-1	-2	-1	2					
Cubic	-1	2	0	-2	1					
Quartic	1	-4	6	-4	1					
6 Levels										
Linear	-5	-3	-1	1	3	5				
Quadratic	5	-1	-4	-4	-1	5				
Cubic	-5	7	4	-4	-7	5				
Quartic	1	-3	2	2	-3	1				
Quintic	-1	5	-10	10	-5	1				
7 Levels										
Linear	-3	-2	-1	0	1	2	3			
Quadratic	5	0	-3	-4	-3	0	5			
Cubic	-1	1	1	0	-1	-1	1			
Quartic	3	-7	1	6	1	-7	3			
Quintic	-1	4	-5	0	5	-4	1			
Sextic	1	-6	15	-20	15	-6	1			
8 Levels										
Linear	-7	-5	-3	-1	1	3	5	7		
Quadratic	7	1	-3	-5	-5	-3	1	7		
Cubic	-7	5	7	3	-3	-7	-5	7		
Quartic	7	-13	-3	9	9	-3	-13	7		
Quintic	-7	23	-17	-15	15	17	-23	7		
Sextic	1	-5	9	-5	-5	9	-5	1		
Septic	-1	7	-21	35	-35	21	-7	1		
9 Levels										
Linear	4	3	2	1	0	-1	-2	-3	-4	
Quadratic	28	7	-8	-17	-20	-17	-8	7	28	
Cubic	-14	7	13	9	0	-9	-13	-7	14	
Quartic	14	-21	-11	9	18	9	-11	-21	14	
Quintic	-4	11	-4	-9	0	9	4	-11	4	
Sextic	4	-17	22	1	-20	1	22	-17	4	
Septic	-1	6	-14	14	0	-14	14	-6	1	
Octic	1	-8	28	-56	70	-56	28	-8	1	
10 Levels										
Linear	-9	-7	-5	-3	-1	1	3	5	7	9
Quadratic	6	2	-1	-3	-4	-4	-3	-1	2	6
Cubic	-42	14	35	31	12	-12	-31	-35	-14	42
Quartic	18	-22	-17	3	18	18	3	-17	-22	18
Quintic	-6	14	-1	-11	-6	6	11	1	-14	6
Sextic	3	-11	10	6	-8	-8	6	10	-11	3
Septic	-9	47	-86	42	56	-56	-42	86	-47	9
Octic	1	-7	20	-28	14	14	-28	20	-7	1
Nonic	-1	9	-36	84	-126	126	-84	36	-9	1

6.2 Appendix B: The Mixed Regression Procedure

Usage:

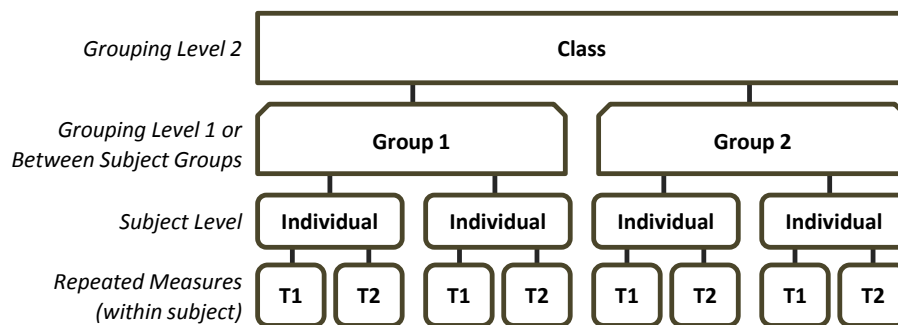
Linear Mixed regression is used when you have **nested** or **repeated data**. **Repeated data** means that the same people are tested multiple time, the data has a correlation and dependency due to the person effect (in a way the repeated measures are nested within person).

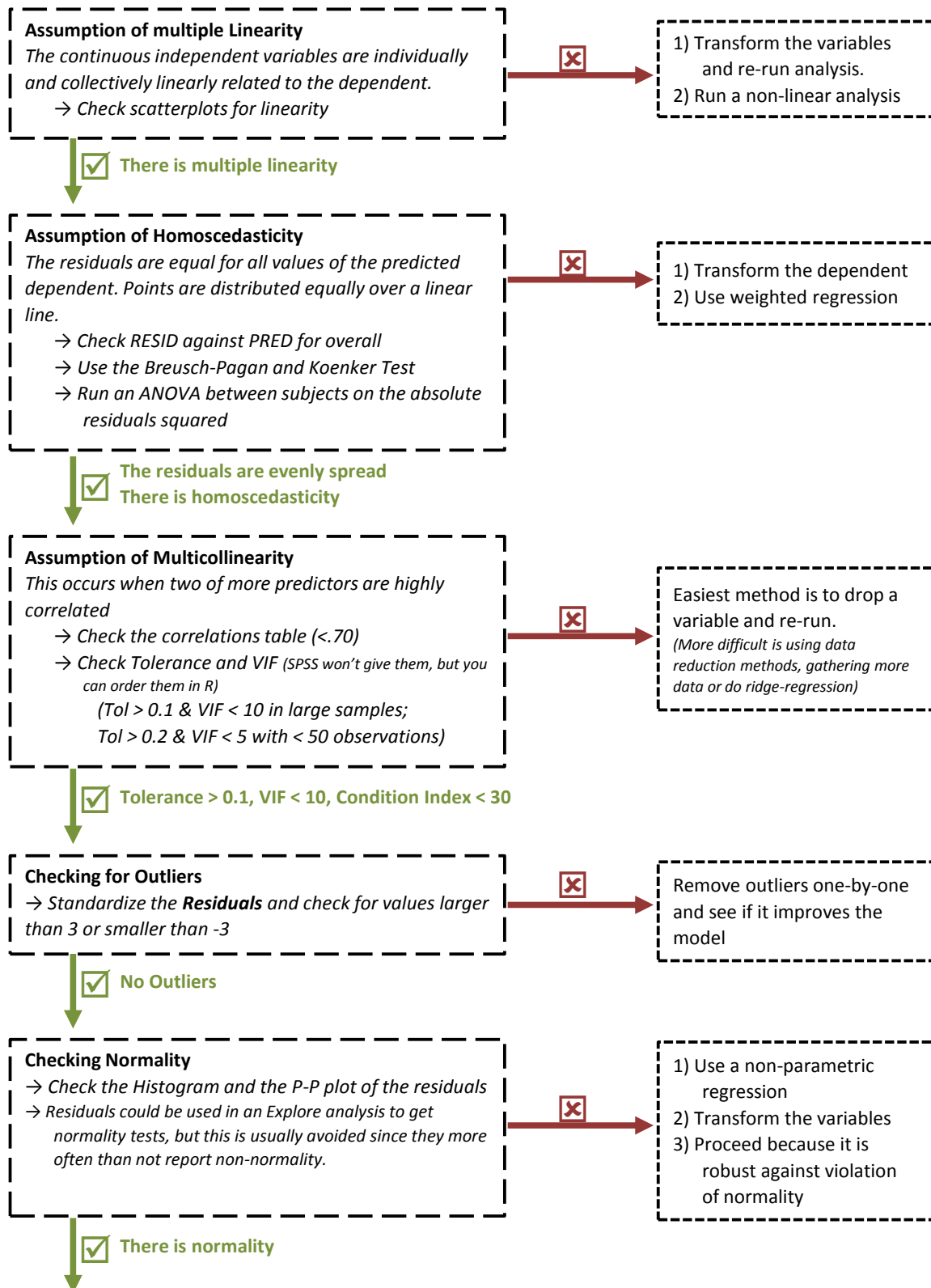
Nested data means that subgroups have a certain degree of similarity. The hierarchy can for instance be classes within schools. Students in the same class will be more similar than students in different classes (level 1), similarly students in the same school will be more similar (level 2).

Other uses for multilevel regression is when you have violated the assumption of independence in regression, when there is missing data, or when you have multiple repeated variables on the same level (like session and order).

Requirements:

- ✓ Two or more independent variable (continuous or categorical) in a hierarchical structure
- ✓ One dependent variable (continuous)
- ✓ Centering is advised for variables that do not have a real-world value of 0 (e.g. weight, height, or fat percentage).
- ✓ When using Mixed Regression for Repeated Measures the data file should be univariate (one row per time-point, giving an x-amount of rows per participant for each time-point)





- Analyze > Mixed Models > Linear
- Start with a fully fixed factorial model (containing all effects and interactions)
 - Add the contextual variables in Subjects
 - These are the variables that group participants. These can be Between Subjects like Schools or Classes. But also Subject Identifiers where you add an ID-variable denoting the person.
 - Add the dependent (to be predicted) variable.
 - Add Predictors
 - Categorical predictors are Factors
 - Continuous predictors are Covariates
 - Fixed: Add fixed effects as Main Effects, Interactions, or polynomials (for WS-factors entered as covariates use the By* function to add Time*Time, and Time*Time*Time to get polynomials)
 - Random: Adding random effects.
 - The Random Slopes are added by adding the factors into the model.
 - Random Intercepts are added by adding the contextual variables (Subject Groupings) to the model (check Include Intercept).
 - Covariance Type: choose the Covariance Matrix.
 - * Variance Components: good enough for random intercept models
 - * Unstructured: good enough for random intercept and random Slope models.
 - * AR(1) Heterogeneous: good enough for WS-designs.
 - Estimation: Check the method used for estimation. ML is used in order to compare models that differ in the fixed part, REML when they differ in the random part. Use REML for the final model for inference.
 - Statistics: Model Statistics (Parameter estimates; Tests for covariance parameters).

```

*Linear Mixed Model Analysis (Syntax dependent on the specified model).
MIXED [Dependent] BY [WS1] [WS2] WITH [Covariate]
/CRITERIA=CIN(95) MXITER(100) MXSTEP(10) SCORING(1)
SINGULAR(0.000000000001) HCONVERGE(0, ABSOLUTE)
LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
/FIXED=[WS1] [WS2] [COV] [WS1]*[WS2] [WS1]*[COV] | SSTYPE(3)
/METHOD=ML
/PRINT=DESCRIPTIVES SOLUTION TESTCOV
/REPEATED=[WS1]*[WS2]*[COV] | SUBJECT([ID Variable]) COVTYPE(CS).

*Optional EMM for Pairwise Comparisons. Interactions can be tested by
specifying a Variable to compare. Covariate values can be fixed at a
value, default is Mean.
/EMMEANS=TABLES([WS1]) COMPARE ADJ(SIDAK)
/EMMEANS=TABLES([WS1]*[WS2]) COMPARE([WS2]) ADJ(SIDAK)
/EMMEANS=TABLES([WS1]*[COV]) WITH ([COV]=MEAN) COMPARE([WS1]) ADJ(SIDAK)
/EMMEANS=TABLES([WS1]*[COV]) WITH ([COV]=[VAL]) COMPARE([WS1])
ADJ(SIDAK)

*Optional Custom Contrasts.
/TEST([Hypothesis value])="Constrast Name" [Variable] -3 -1 1 3

```

Building the Model

These are the important steps in building a model. These can be followed but don't always come in this order. Especially when it comes to model reduction and adding random effects, you might need to go back and forth between changing terms and changing covariance matrices

Comparing Models: Start with a full factorial model

- Check assumptions and remove outliers sequentially (use CS for faster convergence)
 - See if the model makes conceptual sense and whether factors can be combined or have to be split up
- Select the best fitting covariance structure (use REML)
 - Run models using UN, TP, AR1, ARMA, and CS (plus heterogeneous versions)
 - Compare fit using AIC and BIC values, for nested models -2LL can be compared as well
- Reduce the model by removing non-significant terms (use ML)
 - Start with the highest order effects (largest interactions) and remove terms sequentially, check if it improves the fit.
 - Do not remove terms that are part of higher order interactions (keep main effects if they are part of an interaction)
 - Keep effects of interest even if they are non-significant
- Add random effects (use REML)
 - Start with all effects random (all effects you want to check)
 - Remove terms one at a time to see if fit improves
 - You may need to go back and change the covariance matrix (ARMA works quite well) or even add terms back into the model).
- Finalize the model (use REML)
 - Set the final model to REML for inference