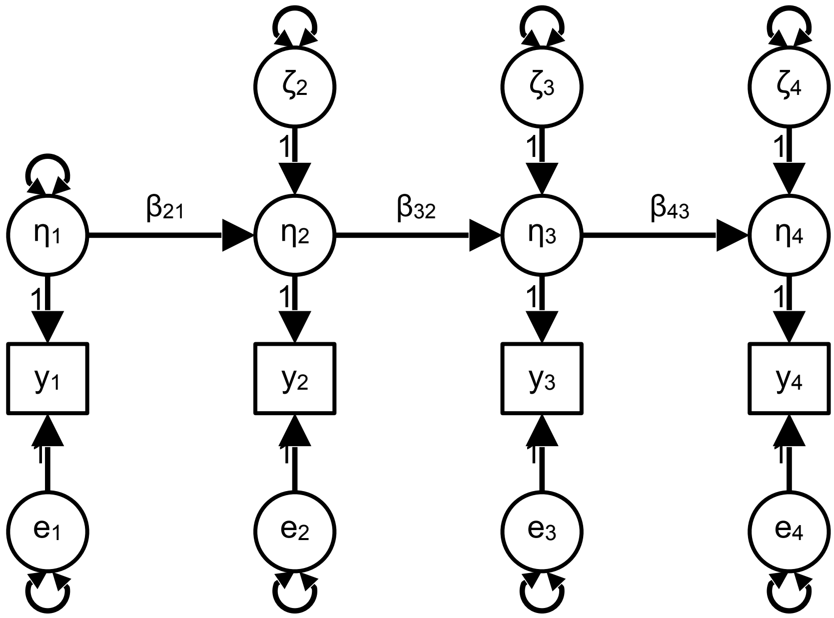
Homework for lab 3

Exercise 1

The covariance matrix is in the file Coenders.dat comes from 1724 children and adolescents that participated in the National Survey of Child and Adolescent Well-Being (NSCAW) in Russia. They indicated how satisfied they were with their lives as a whole on a 10-point scale (1=not at all satisfied, 10=very satisfied). There were three waves (1993, 1994 and 1995). At the third wave, the question was asked twice (with 40 minutes in between). Hence, in total there are four measurements obtained at three waves.

The researchers are interested in fitting a quasi-simplex model to these data, that is, a simplex model at the latent level, thus accounting for measurement error in the observations. Below a graphical representation of this model is given.



a. Provide the names of the variances (i.e., indicate in which model matrix, and which position in this matrix they have) in the graph above. What is the difference between the e's and the ζ's?
A: The e's are part of the measurement equation, whereas the Dzeta's are part of the structural equation. This means that the dzeta's are the (residual) variances (also known as dynamic error) (after regression the variables on one another, which are in Psi) and the e's are the measurement errors (which are in Theta).

b. How would you specify the model in Mplus?
A: In the following way:

```
TITLE:  Exercise 1
DATA:   File is Coenders.dat;
        TYPE is COVARIANCE;
        NOBS is 1724;
VARIABLE: NAMES = t1 t2 t3a t3b;
MODEL:
        f1 by t1@1;
        f2 by t2@1;
        f3 by t3a@1;
        f4 by t3b@1;

        f2 ON f1;
        f3 ON f2;
        f4 ON f3;

OUTPUT: SAMPSTAT TECH1;
```

c. Determine the df for this model (indicate how you obtained this number). Is it possible to estimate this model?
A: Sample statistics = elements in covariance matrix = 4 * 5 / 2 = 10
Free parameters = 3 regression coefficients, 4 (residual) factor variances, 4 residual variances = 11
➔ Df = 10 – 11 = -1… So no it is not possible to estimate this model currently.

d. To make sure a quasi simplex model is identified, often the variances of the measurement errors are constrained to be equal over time. How can you do this in Mplus? How many df does this model have?
A: This means the e's have to be equal: we do this in the following way

```
MODEL:
         f1 by t1@1;
         f2 by t2@1;
         f3 by t3a@1;
         f4 by t3b@1;
         t1 - t3b (resvar);
         f2 ON f1;
         f3 ON f2;
         f4 ON f3;
```

➔ Now, the model has 8 free parameters (3 regression coefficients, 3 residual factor variances, 1 factor variance and 1 measurement variance (assumed equal for 4 items))
df = 10 – 8 = 2

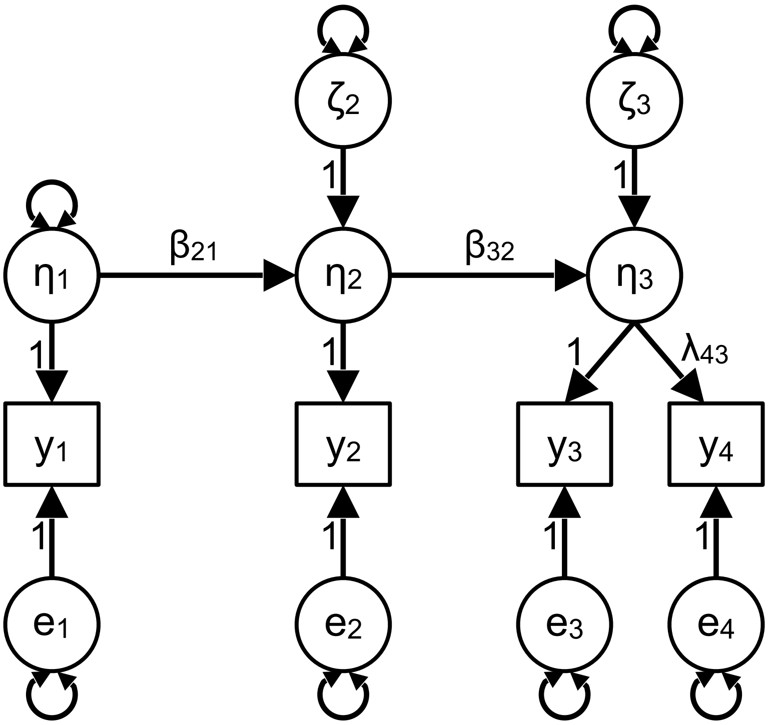
e. Run the model and report on the model fit.
A: Goodness of fit test showed the following results:
chi-square = 13.29, df = 2 (significant, bad fit, but a lot of power)
RMSEA = 0.057 (good fit)
CFI/TLI = 0.994 / 0.981 (good fit)
SRMR = 0.025 (good fit)
➔ The model fits good.

f. The quasi-simplex model you just ran, led to the following warning:

```
WARNING:  THE LATENT VARIABLE COVARIANCE MATRIX (PSI) IS NOT POSITIVE
DEFINITE.  THIS COULD INDICATE A NEGATIVE VARIANCE/RESIDUAL VARIANCE FOR A
LATENT VARIABLE, A CORRELATION GREATER OR EQUAL TO ONE BETWEEN TWO LATENT
VARIABLES, OR A LINEAR DEPENDENCY AMONG MORE THAN TWO LATENT VARIABLES.
CHECK THE TECH4 OUTPUT FOR MORE INFORMATION.
PROBLEM INVOLVING VARIABLE ETA4.
```

What is the problem?
A: There seem to be a significantly negative residual variances in Psi, namely for eta4.

As indicated in the description of the data, the third and fourth measurement were obtained at the same measurement wave (with only 40 minutes in between). Hence, the researchers proposed the following model instead of the regular quasi-simplex model.



g. Explain why this model makes more sense for these data than the regular quasi-simplex model.
A: This model makes more sense, because the last measurement is measured twice. It does not make sense that within 40 minutes, a whole new latent variable is responsible for life satisfaction.

h. How many df does this model have? Note that we keep the constraint on the variances of the measurement errors.
A: Free parameters = 2 regression coefficients, 1 factor loading, 2 residual variances, 1 factor

variance, 1 variance constrained equal = 7
➔ Df = 10 – 7 = 3

i. Are these two models nested? If so, how? If not, why not, and how could we compare them?
A: It turns out… these models are nested. This is because the variance of eta4 was not really there anyway (as it was significantly negative, and hence probably zero) and therefore nothing was added by adding the fourth latent variable. Hence, they are nested and we can use a chi-square difference test.

j. Specify this model in Mplus and run it. Report on the model fit.
A: MODEL:

```
f1 by t1@1;
f2 by t2@1;
f3 by t3a@1 t3b;
t1 - t3b (resvar);
f2 ON f1;
f3 ON f2;
```

Model fit:
chi = 27.37, df = 3 (significant, bad, but power)
RMSEA = 0.069 (good)
CFI/TLI = 0.987 / 0.973 (good)
SRMR = 0.039 (good)
→ Still fits good

k. Compare the two models to each other: What can you conclude?
A: With a chi-square difference test, we find that it does lead to a significant loss of model fit. The AIC’s also differ by 7, which is quite large. But considering that we no longer get an error (or a significantly negative variance) and the fit indices are still good. I would state the second model is better and should still be used.

l. Can you improve the second model in any way? Indicate which parameter you would add to your model, and what this parameter represents in substantive terms.
A: According to the modification indices, we can do many things that would significantly improve model fit. A few are: freeing a measurement variance, allowing a correlation between measurement variances, adding another regression coefficient (from eta 1 to eta3), allowing items to load on multiple factors, etc.

m. For the second model, write down the covariance between Y3 and Y4 and between Y1 and Y3 in terms of model parameters.
A: I’m good thank you.

Exercise 2

In this exercise we make use of the GPA data that was also used in the multilevel course. The data can be found in GPA.dat, and contain:

- student number
- sex (0=male ; 1=female)
- high school GPA
- GPA1-GPA6 (GPA)
- job1-job6 (number of hours spend on job)
- admitted to a university (0=no; 1=yes).

You can select the GPA-scores with the USEVARIABLES command in Mplus (check the Mplus User's Guide).

a. We are interested in fitting a quadratic LGC model on the GPA scores. Indicate the number of observed statistics, the number of parameters in the model, and the df.
A: Sample statistics = 6 + 6 * 7 / 2 = 27
Free parameters = 3 variances (i.e., 3 random effects), 3 covariances (between random effects), 3 means (fixed effects), 6 residual variances = 15
→ DF = 27 – 15 = 12

b. Specify the quadratic LGC model for the GPA scores, using:
MODEL: i s q | GPA1@0 GPA2@1 GPA3@2 GPA4@3 GPA5@4 GPA6@5;
Does the model fit? Discuss the fit measures.
A: chi-square = 9.86 (df = 12) (non-significant, good)
RMSEA < .001 (good)
CFI/TLI = 1 (good)
SRMR = 0.029 (good)

c. Next, consider using the command:
MODEL: i s q | GPA1@-5 GPA2@-3 GPA3@-1 GPA4@1 GPA5@3 GPA6@5;
How does this differ from the previous specification?
A: The results are more or less the same. However, the intercept point is at 0, so this is now in the exact middle (i.e., we centered)

d. Run the second model, and report on the fit.
A: The model fit is exactly the same as for the first specification (i.e., they are statistically equivalent).

e. Which parameters (in each model) describe the average growth curve? What can you say about the shape of the average growth curve based on the parameter values?
A: The average growth curve is determined by the fixed effects. In Mplus this entails the mean of the factors! These are: 2.862, 0.053, 0.000. From these we can state that the average growth curve is linearly increasing (as the fixed quadratic effect is not significant (but the random effect is))

f. Include the statement RES on the output line and discuss the additional output you get: What can you use this for?
A: Now, we get the residual output. This gives you the estimated model and residuals (observed and estimated – also standardized). You can use these to get the model implied covariances/means, etc.

g. Describe the individual differences in growth curves in both models.
A: We find that there are significant individual differences for both the intercept, the linear trend and the quadratic trend as the variance of I, S and Q is significant. So although on average there is no quadratic effect, individuals do have a quadratic effect that has a variance.

h. Consider the covariances between the random effects in both models: What can you conclude from them?
A: We find that there is a significant covariance between I and S andQ. This means that the intercept variance depends on the slope variance, etc. Hence it is important to take this into account when interpreting the results (it becomes very hard…).

i. Note that the residual variances are unconstrained over time, as opposed to doing this kind of analysis in HLM of with lme() in R (then there is just one residual variance at level 1). Can you think of a way to test whether the residual variance is constant over time, or that it differs? Describe your approach, and the results.
A: Perhaps simply adding a new parameter that is the difference between two residual variance and seeing whether this is significant. Or constraining them to be equal and doing a chi-square difference test to see if it leads to a significant loss of model fit.

j. Suppose you would use sex and high school GPA as covariates (=predictors) of the random intercept, slope and quadratic term. What does it mean if these prove to be significant covariates? Can we compare this model to the model without these level 2 predictors?
A: If these prove to be significant, it means we are explaining variance in the intercept, slope and quadratic term variance. We can not compare this model to the previous model however, because they are based on a different covariance matrix (so we can't even use AIC / BIC).

k. Is there a problem with sex being a dummy variable?
A: A dummy variable is fine, as long as it is an exogenous variable. Otherwise, some stuff would have to be changed.

Exercise 3

In the paper by Hamaker, Kuiper and Grasman (2015), an alternative to the traditional cross-lagged panel model is proposed, which can be used to separate the dynamics of the within-person process from stable, between-person differences. This is illustrated with bivariate data measured at three waves.

a. Check the Mplus output files *clpmpc1.out* and *clpmpc1a.out*. Describe the models that are fitted and how they differ from each other.
   A: There are two variables that are measured at three occasions. Furthermore, everything is allowed to covary with everything (i.e., it is a saturated covariance matrix - only the means are constrained).
   Difference:
   In the first model, they set equal the means of the three different timed PSCon and DEP variables, whereas in the second model the first measured DEP mean was estimated freely.

   Turns out, you can constrain the means except for first DEP mean (as this had a M.I. of 10). That is why the second model is used.

b. Check the Mplus output file *clpmpc2a.out*, and describe the model that is estimated (TIP: write down what the different parts of the model command are doing.)
A: Like second model, first mean of DEP can be freely estimated again, other variables have means that are set equal to their counterparts. Measurement error variance is set to 0 for all variables. And instead of having covariances everywhere, they are now regression later variables on earlier variables (but also allowing them to covary DEP and PScon at same time)
The latent variables are added (compared to using only the items) is because Mplus will otherwise do this by default and then some alpha's (which are set equal) will be means and intercepts… which you do not want to be equal!

c. Check the Mplus output file *clpmpc3a.out*, and describe the model that is estimated there.
A: Now the variances of the continuous latent variable counterparts have variances set to 1, but measurement error variance still set to 0. Furthermore, the means of first is set free again are now made to be latent variables and are not allowed to covary with anything else. The means of the variables however, are now turned into a latent construct (MD and MPs) to separate the within and between part of the model.

The latent mean structure now stays stable over time (so it is variation <u>between</u> person) and the other structures (that are being regressed on one another) are the <u>within</u> variation!

d. How is are the last two models related to each other, and what can you conclude about these models?
A: These models are different in the sense that the second model (in c) now separates the within and between variation. The model in B is nested under model C.

Exercise 4

a. For Exercise 1, Lab 1, give a brief description of the model (e.g., 1-factor model; path analysis; multi-group factor analysis), and summarize the most important learning objectives.
MODEL: two factor model for one group
we looked at defaults in Mplus, model matrices and how to fix parameters

b. For Exercise 2, Lab 1, give a brief description of the model (e.g., 1-factor model; path analysis; multi-group factor analysis), and summarize the most important learning objectives.
MODEL: two factor model
getting data from SPSS in Mplus, write the model in terms of equations, different ways of scaling and we saw how the one factor model is nested under two factor model

c. For Exercise 3, Lab 1, give a brief description of the model (e.g., 1-factor model; path analysis; multi-group factor analysis), and summarize the most important learning objectives.
MODEL: second order factor model (intelligence)
This is done by using matrix Beta, looked at non-significant residual variances, nesting of second order factors

d. For the first Hand-in assignment, give a brief description of the model (e.g., 1-factor model; path analysis; multi-group factor analysis), and summarize the most important learning objectives.
MODEL: two factor model
ON statement for latent variables, modification indices, R^2

e. For Exercise 1, Lab 2, give a brief description of the model (e.g., 1-factor model; path analysis; multi-group factor analysis), and summarize the most important learning objectives.
MODEL: multiple group factor analysis
different types of measurement invariance (configural, weak and strong), defaults, TECH1, overruling defaults. Bias (through MI) and interpretation.

f. For Exercise 2, Lab 2, give a brief description of the model (e.g., 1-factor model; path analysis; multi-group factor analysis), and summarize the most important learning objectives.
MODEL: (path) mediation analysis
model matrices in this case, defaults (covariances between residuals of endogenous variables, exogenous variables no means/variances), indirect/total effect, comparing effects with standardised results

g. For Exercise 3, Lab 2, give a brief description of the model (e.g., 1-factor model; path analysis; multi-group factor analysis), and summarize the most important learning objectives.
MODEL: multiple group two factor model
looked at negative variances (Heywood case), strong-factorial invariance, constraining parameters across groups, regression with latent variables

h. For Exercise 4, Lab 2, give a brief description of the model (e.g., 1-factor model; path analysis; multi-group factor analysis), and summarize the most important learning objectives.
MODEL: multiple group factor analysis on simulated data
Showing that no matter which item is used for scaling (biased or not) does not matter at all.

i. For the second hand-in assignment, give a brief description of the model (e.g., 1-factor model; path analysis; multi-group factor analysis), and summarize the most important learning objectives.
MODEL: two mediation models (one nested under other)
Translating hypotheses into models

j. For Exercise 1, Lab 3, give a brief description of the model (e.g., 1-factor model; path analysis; multi-group factor analysis), and summarize the most important learning objectives.
MODEL: pass simplex model
going from non-identified model to Heywood case to a final model, looking at nesting, sometimes worse fit models must still be kept (if negative var otherwise)

k. For Exercise 2, Lab 3, give a brief description of the model (e.g., 1-factor model; path analysis; multi-group factor analysis), and summarize the most important learning objectives.
MODEL: latent growth curve modelling with linear and quadratic slope
Fixed effect does not have to be significant for there to be a random effect, RES output to see average growth curve, interpretation = complicated

l. For Exercise 3, Lab 3, give a brief description of the model (e.g., 1-factor model; path analysis; multi-group factor analysis), and summarize the most important learning objectives.
MODEL: Cross-lagged panel models

Seeing what constraining means can be done and separating mean structure through latent variable to divide between and within variation.