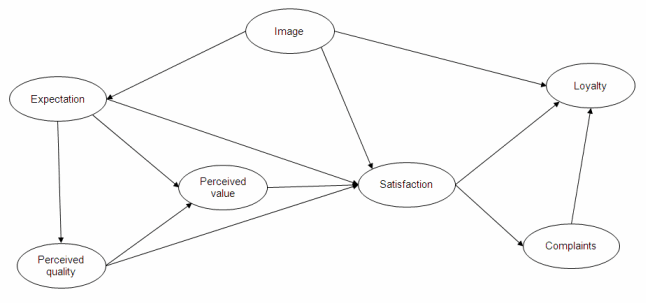
**Statistics: Structural Equation Modeling**

**Lecture 1 An Introduction to SEM: Specification, Estimation, Testing and Missingness**

In traditional statistics you explain difference sin scores or means. Each statistical procedure has its own estimation formula and test 🡪 but you always try to predict scores. For example a t-test tries to test for differences in means and an ANOVA tests the effect of an intervention. In the end it’s all about prediction.

But in structural equation modeling you try to explain relationships between variables. You look at the correlation between variables. SEM can have more than one dependent variables. You try to explain direct and indirect relationships 🡪 how are variables related and how are they not related. So you look at the relationship between variables through other variables. Relationships are everywhere.

Path model 🡪 visualization of a set of hypotheses which together form a theory:



There can be many dependent variables, which can be independent variables at the same time. SEM tries to explain relationships between variables in terms of model parameters. SEM is a method to visualize, estimate and test a network of relations between variables. The arrows are hypotheses and the model as a whole represents a theory. This theory can explain:

* Correlations between variables
* Variances of the variables
* Means of the variables

ANOVA has a closed form of expression 🡪 you feed the data into a formula and you will find a single unique solution for each variable. But when you want to extend the model, you cannot do this with ANOVA. My precious:

∑ = ∑ (θ)

(θ) 🡪 model parameters

∑ (θ) 🡪 model implied correlation

There are rules behind this formula that transform the model into correlation. Each model leads to different sigma’s and theta’s. so when you change the model you change the way the correlations are calculated. These rules are:

* The correlation between two variables is equal to the sum of the direct effect, the indirect effects, confounding effects and joint effects between these variables
* The total variance of an endogenous variable is equal to the sum of the amount of variance explained by the causal variables of this endogenous variable and the amount of unexplained variance.

My precious enables finding the population values for the model parameters. You can get the estimations form the model when you work out the correlations in combination with beta and theta.

My precious only works for ∑ 🡪 the population correlation matrix. But we always analyze the sample correlation matrix S, however S is not equal to ∑. That’s why we need estimators that are:

* Consistent. When the sample size gets larger then the probability to find the population value must also get larger
* Unbiased. The expected value of the estimator over many samples of the same size is the population value
* Efficient. The variance of the estimator, or the standard error, is as small as possible

In SEM there are many estimators available. With the SEM analysis you try to estimate the mean and the variance of the parameters. By choosing different weights you can get different estimators with different characteristics. One estimator is ML, maximum likelihood 🡪 is consistent, unbiased and efficient. Another one is ULS, unweighted least squares 🡪 is consistent and unbiased but not efficient. ML is the default in SEM.

ULS 🡪 W = I

ML 🡪 W = sifma hat.png

Lavaan code is used to estimate the model.

Z value 🡪 can tell you if a parameter is significant. If the Z value is bigger than one, then the parameter is significant. They you’re hypotheses are confirmed.

ML performs quite good under general conditions:

* A sample size between 200 and 500 are usually sufficient
* You need 10 cases per parameter at a minimum

Sample size requirements increase when you have a smaller R2, when collinearity is greater, for smaller numbers of indicators per factor and when the data is not normal. Nonnormality is a major threat to accurate estimation. Solutions against nonnormality are available 🡪 there are special robust estimators. However, ML isn’t that bad as was found in robustness studies.

Nonnormality is no problem if the skewness is smaller than 2 and if kurtosis is smaller than 7. Also categorical data with as little as 3 categories is acceptable if the distribution is normal. There’s not rule of thumb that applies to all situations! But in general LM is ok!

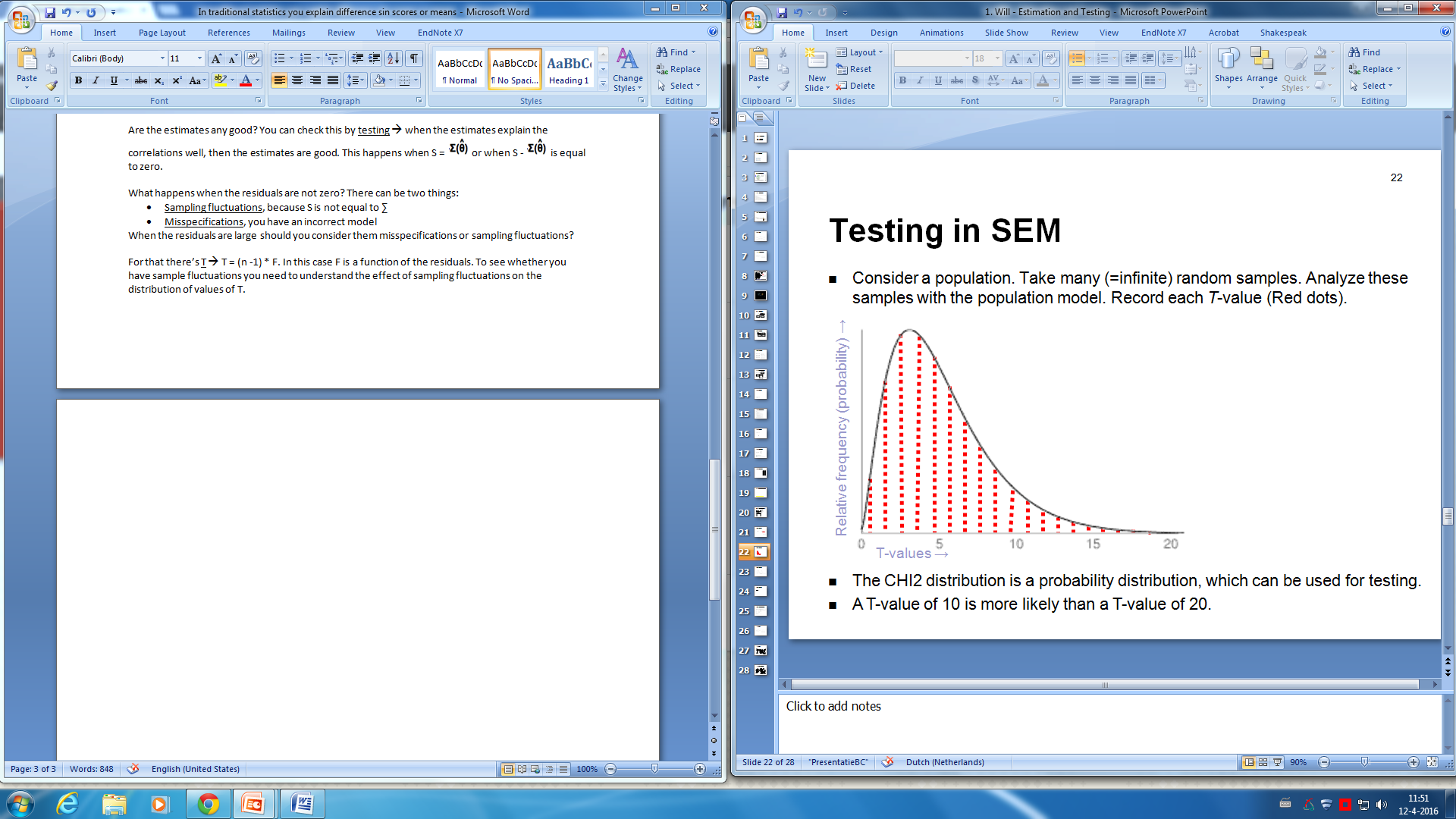
Are the estimates any good? You can check this by testing 🡪 when the estimates explain the correlations well, then the estimates are good. This happens when S = sifma hat.png or when S - sifma hat.png is equal to zero.

What happens when the residuals are not zero? There can be two things:

* Sampling fluctuations, because S is not equal to ∑
* Misspecifications, you have an incorrect model

When the residuals are large should you consider them misspecifications or sampling fluctuations?

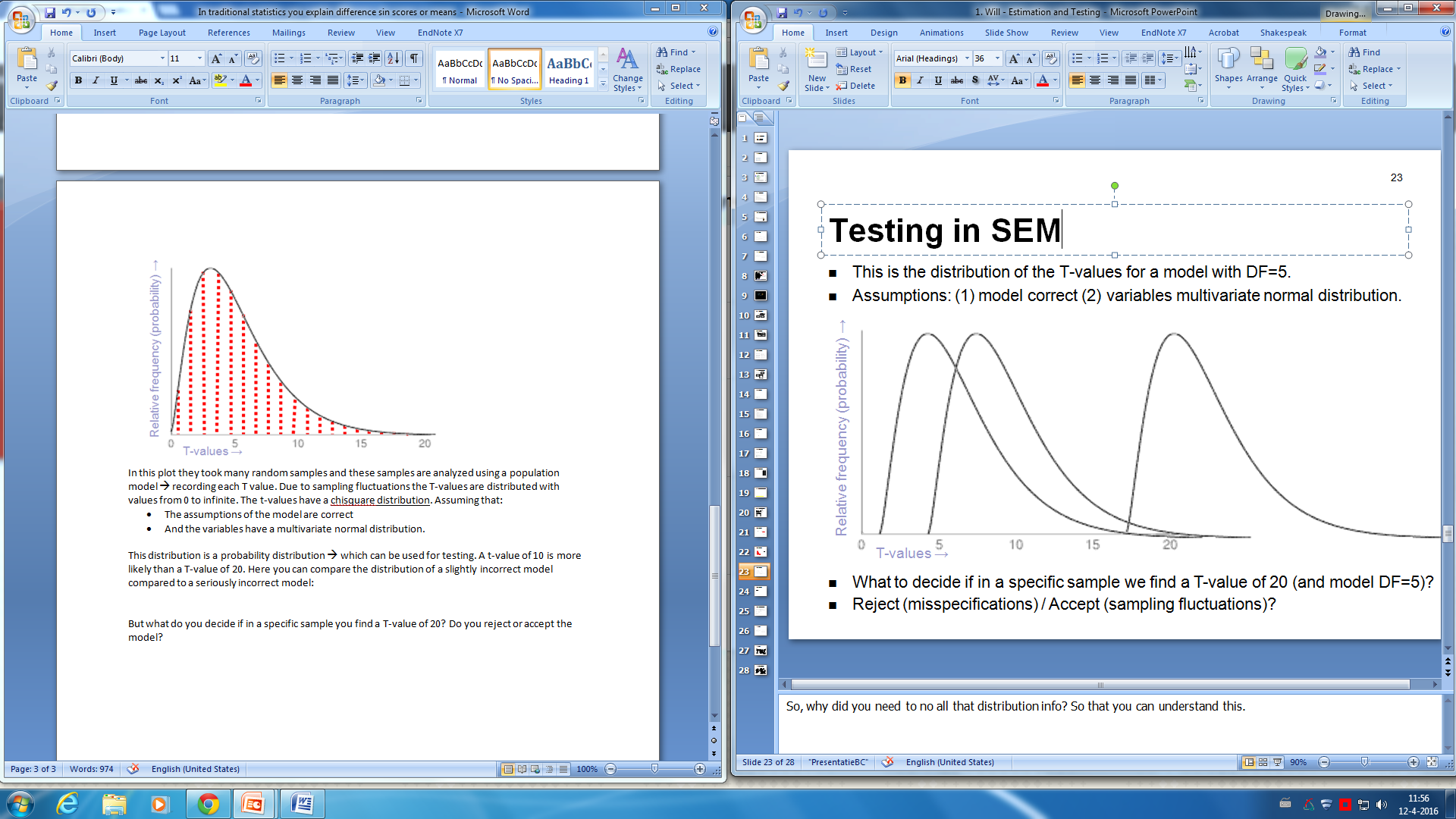
For that there’s T 🡪 T = (n -1) \* F. In this case F is a function of the residuals. To see whether you have sample fluctuations you need to understand the effect of sampling fluctuations on the distribution of values of T.



In this plot they took many random samples and these samples are analyzed using a population model 🡪 recording each T value. Due to sampling fluctuations the T-values are distributed with values from 0 to infinite. The t-values have a chisquare distribution. Assuming that:

* The assumptions of the model are correct
* And the variables have a multivariate normal distribution.

This distribution is a probability distribution 🡪 which can be used for testing. A t-value of 10 is more likely than a T-value of 20. Here you can compare the distribution of a slightly incorrect model compared to a seriously incorrect model. A completely incorrect model shifts to the right:



But what do you decide if in a specific sample you find a T-value of 20? Do you reject or accept the model? You always have the risk of rejecting a good model or accepting a wrong model. Where do you draw the line? What risk are we willing to take to reject a correct model? In general 5% risk is seen as acceptable in social sciences, this risk level is called the significance level of the test. Therefore, when the p value is below 0.05, the model is judged correct.

To check whether this is the case you first have to look at the model and then at the estimates. You should not evaluate the parameters, the hypotheses, if the model is rejected.

SEM is very flexible 🡪 you can specify any model in the same framework. So you can also use traditional statistics in SEM, like ANOVA or regression. You can choose which parameters to estimate and which not.

Latent variables 🡪 should be represented in circles

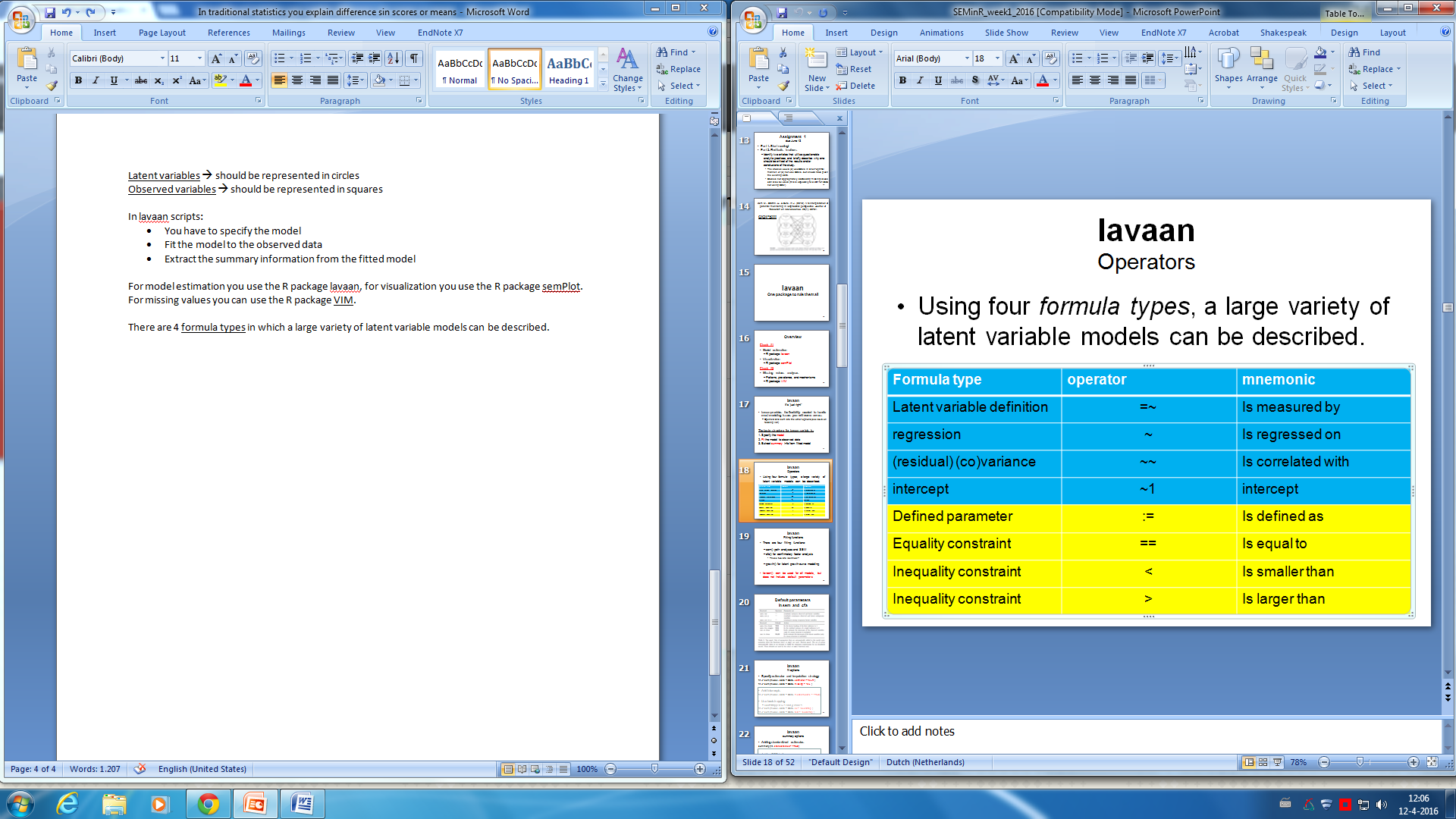
Observed variables 🡪 should be represented in squares

In lavaan scripts:

* You have to specify the model
* Fit the model to the observed data
* Extract the summary information from the fitted model

For model estimation you use the R package lavaan, for visualization you use the R package semPlot. For missing values you can use the R package VIM.

There are 4 formula types in which a large variety of latent variable models can be described. Also, there are different operators for lavaan.



There are four fitting functions:

* Sem(), which uses path analyses and SEM
* Cfa(), is for confirmatory factor analysis
* Growth(), is for latent growth curve modeling
* Lavaan(), can be used for all models but does not include default parameters. So you have to know exactly what you are doing. While the other three commands have a lot of details, are easier because they have more defaults.

In lavaan you have to specify the estimator and the imputation strategy by using the command. Put this in the sem command 🡪 fit < - sem(model, data = Data, estimator = ‘MLR’, missing = ‘ML’)

If you want to add intercepts use meanstructure = TRUE. And when you want to do bootstrapping:

* fit < - sem(model, data = Data, se = ‘bootstrap’) 🡪 for standard erros
* fit < - sem(model, data = Data, test = ‘bootstrap’) 🡪 for p-values

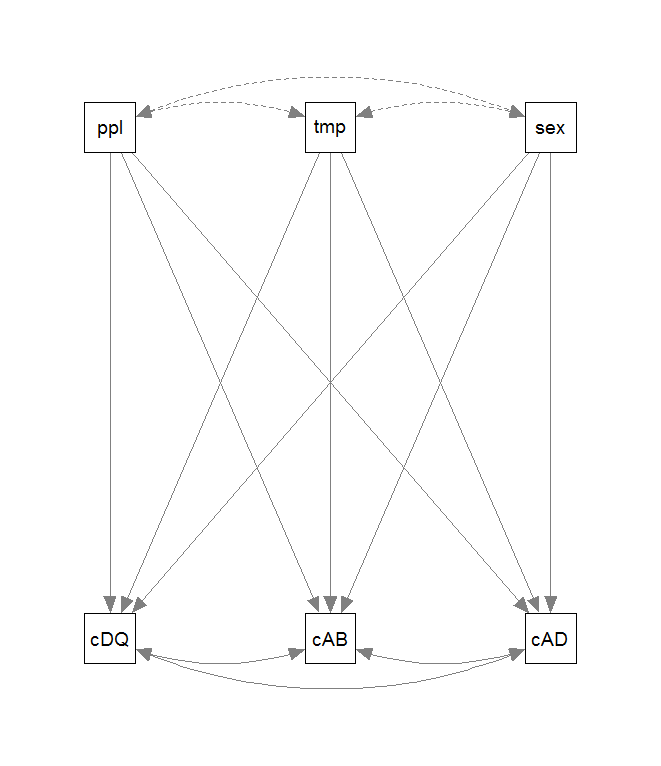
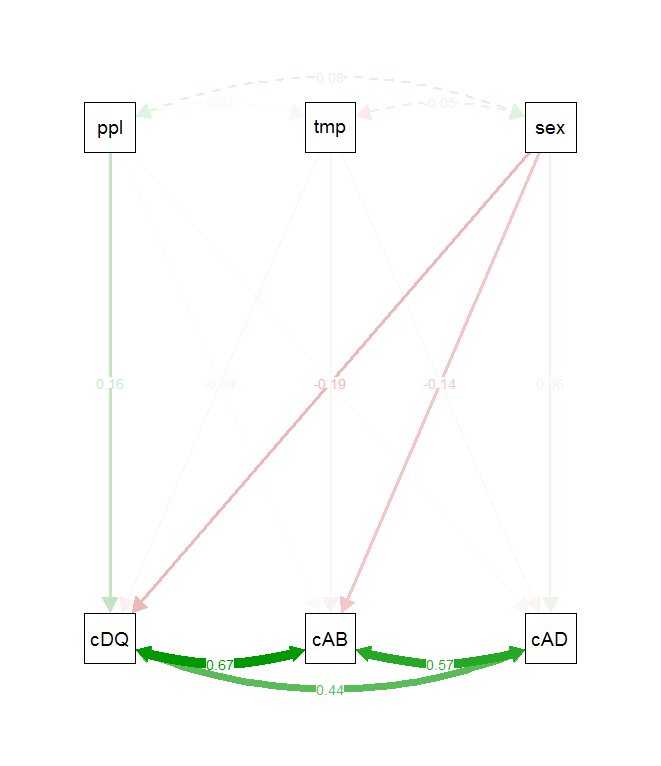
then you can look at the summary of this fit. You can add standardized estimates by using summary(fit, standardized = TRUE) and you can add GOF indices by using summary(fit, fit.measures = TRUE). And lastly if you want to add modification indices use summary(fit, modindices = TRUE).

When you look at the output, look at the std.all instead of the std.lv. These are you’re betas. If all endogenous measures are continuous, the default estimator is maximum likelihood 🡪 for both complete and incomplete data. But if they are not you can switch to another estimator.

Alternatives to ML are:

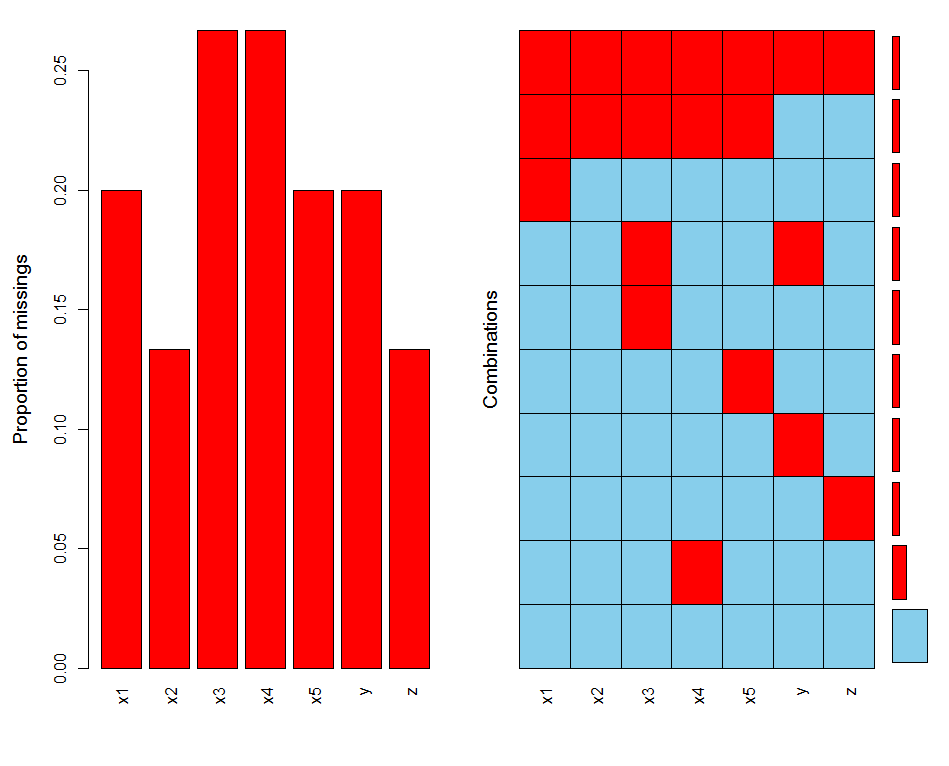
* GLS. Generalized least squares, which is for complete data only
* WLS. Weighted least squares, which is for complete data only
* DWLS. Diagonally weighted least squares
* ULS. Unweighted least squares

With semPlot you can visualize SEM 🡪 use the function semPaths(). This package uses lavaan based objects and you can also look at significant paths.

semPaths(fit, residuals = F) 🡪 did you specify your model correctly?

Missingness can be on three levels:

* Items. Leaving an item or two balnk form multiple items scales
* Scales. Omitting answers for an entire scale or construct
* Entire surveys. Individuals fail to complete an entire survey. The prevalence of missingness can be assessed at each of these levels. How many variables, participants and values/cells have missing values? The focus should usually be on cases/participants.

In the left graph you can see the proportion if missingness per variable. And in the right graph you can see the missingness on each variable per participant, the proportion of participants that have missing values on a given variable. So you can see that 1 participant misses all variables.

Problems caused by missingness:

* Questionable external validity 🡪 response rate bias. Which means that results from the sample may not be identical to those obtained from the entire sample
* Low statistical power 🡪 the sample is too small to yield a statistically significant results. Which leads to type 2 errors. Note, only smaller samples result in low statistical power, not missingness per se.
* Potentially invalid conclusions!

But why are your data missing? What are the underlying mechanisms? The missing values can be ignorable:

* MCAR, missing completely at random. The probability of missingness does not depend on the observed or missing values
* MAR, missing at random. The probability of missingness partly depends on the observed values, but not on the missing values. Missingness is systematic!

They can also be non-ignorable 🡪 MNAR, missing not at random. Which means that the probability of missingness depends on the missing values themselves.

The bad news is that you cannot definitively determine whether missingness is MCAR, MAR or NMAR. You cannot determine if missingness on Y depends on the missing values of Y 🡪 they are missing! NMAR is based on partially untestable assumptions, and MCAR can never be proven 🡪 only disproven!

The good news is that you can determine if your data are not MCAR and if listwise deletion is biased! You can do this by using MCAR test from the package BaylorEdPsych. Under MCAR the calculated means of the observed data should be the same irrespective of the pattern of missingness. The null hypothesis is that the data are MCAR. If the data are not MCAR 🡪 the mean scores at each assessment will vary across the patterns.

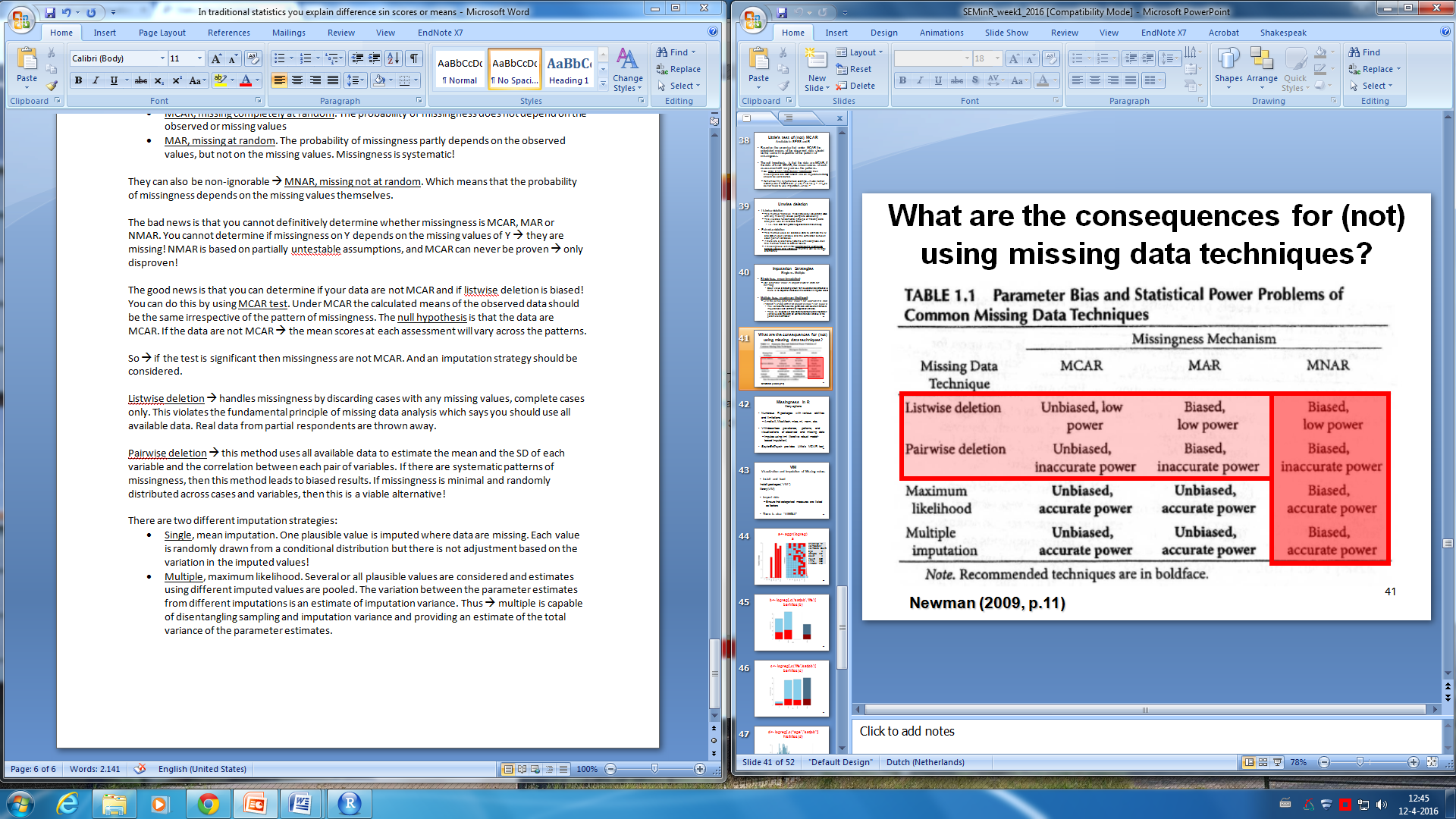
So 🡪 if the test is significant then missingness are not MCAR. And an imputation strategy should be considered.

Listwise deletion 🡪 handles missingness by discarding cases with any missing values, complete cases only. This violates the fundamental principle of missing data analysis which says you should use all available data. Real data from partial respondents are thrown away.

Pairwise deletion 🡪 this method uses all available data to estimate the mean and the SD of each variable and the correlation between each pair of variables. If there are systematic patterns of missingness, then this method leads to biased results. If missingness is minimal and randomly distributed across cases and variables, then this is a viable alternative!

There are two different imputation strategies:

* Single, mean imputation. One plausible value is imputed where data are missing. Each value is randomly drawn from a conditional distribution but there is not adjustment based on the variation in the imputed values!
* Multiple, maximum likelihood. Several or all plausible values are considered and estimates using different imputed values are pooled. The variation between the parameter estimates from different imputations is an estimate of imputation variance. Thus 🡪 multiple is capable of disentangling sampling and imputation variance and providing an estimate of the total variance of the parameter estimates.



There are numerous R packages with various abilities and limitations. VIM describes the prevalence, patterns and visualizations of observed and missing data. When you use this package make sure that categorical measures are listed as factors.

**Lecture 2 Model Improvement**

Maximum Likelihood is the default in SEM. ML is asymptotically consistent, efficient and unbiased. But under which assumptions do I get the correct estimation?

* The observed variables are multivariate normally distributed. In structural models this should be the case for only the endogenous variables. In factor models this should be the case for all indicators. If this assumption is violated, the estimates are still consistent but not efficient any longer.
* The observed variables are observed without measurement error. In structural models this should be the case for all variables. In factor models indicators may be observed with error. But if this assumption is violated, estimates are biased, not consistent and not efficient any longer. A possible solution could be to correct for measurement error using CFA.

If the model is not correct you cannot trust the estimates and you cannot draw conclusions from the model. Model evaluation is essential to test whether there is empirical support for the model, the idea. A model is a simplified idea about reality. But if p < .05 🡪 you have to reject the model.

ML requires complete data, missingness is a big problem! If you have MNAR there’s nothing you can do, except change the population. In the case of MCAR of MAR you can use:

* Listwise deletion. Is okay but you will lose efficiency
* Single imputation. Is unwise because the correlation matrix often is not positive definite.
* Multiple imputation. Good, but yields different results with every analysis
* Full information maximum likelihood, FIML. Is the best to analyze the raw data

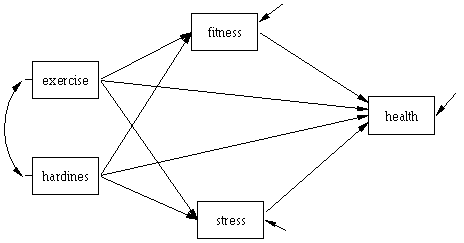
How can you improve the model if it doesn’t fit?

The degrees of freedom of the chisquare test is not the same for all models. In contrast to traditional statistics the df is not dependent on the sample size. But it is dependent on the number of observed variables in the model (A) and the number of parameters that have to be estimated (B).

Df = A – B

A = n (n+1) / 2

B = estimated parameters in a model

For example in this case:

A = 5 (5+1) / 2 = 15

B = 14

Df = 15 – 14 = 1

There can be three cases:

* If df < 0 🡪 you cannot come to a single solution. There is no solution at all. These models are called not identified.
* If df = 0 🡪 there’s a single unique solution for each parameter. So this always leads to a perfect solution. Chisquare is 0, p-value is 0 and the residuals are zero. Models like these are called saturated models. But you cannot test how well correlations between variables are explained. However, you can test how well an x-variable explains a y-variable. So you can test the model but not the parameters.
* If df > 0 🡪 there is only 1 best solution for each parameters. This best solution is found by minimizing F. A test for the model is available so typically P is smaller than 1 and the residuals are not 0. You can test how well correlations between variables are explained and you can test how well an x-variable explains a y-variable, but only if the p-value is bigger than 0.05.

Even when df >= 0 a model might not be identified because the sufficient conditions for identification are not fulfilled. Traditional statistical models always have df = 0 in their SEM equivalents.

A model is misspecified when:

* Parameters that you estimate are actually zero in the population 🡪 over parameterization
* Parameters that you do not estimate (fixed to zero) are actually not zero in the population (under parameterization)
* Both

Then you can look at the modification index 🡪 you estimate the parameters and they are fixed. Then you estimate a misspecified parameter. This will result in:

* A value for that parameter 🡪 this value is called the expected parameter change, EPC.
* A decrease in the chisquare value, this decrease is called the modification index. If this decrease is significant (MI > 3.84), then this indicates that the introduction of that parameter will lead to a significant change in the fit of your model.

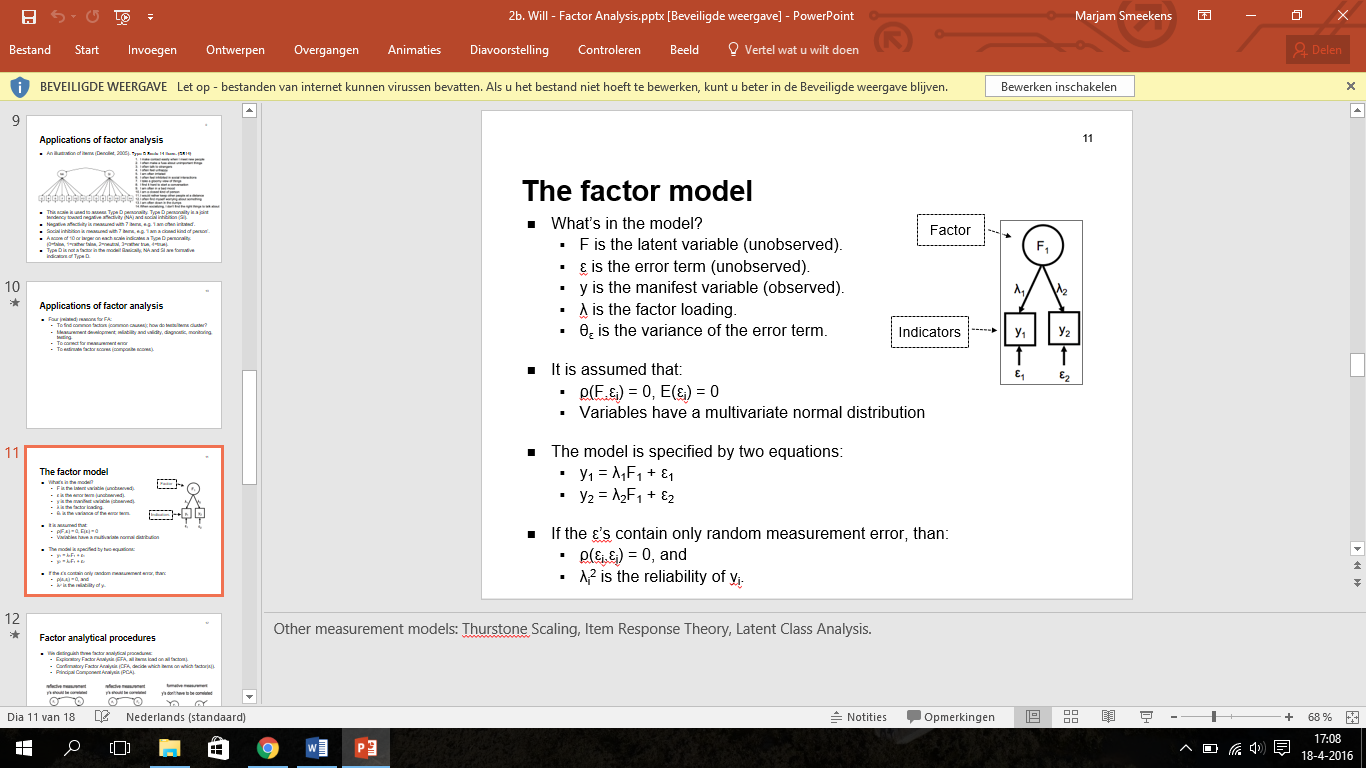
Estimates cannot be trusted if p < 0.05!

Factor analysis 🡪 a factor means a doer or an agent, it’s something that makes something else happen. Factor analysis originated from ideas from Darwin, Galton and Spearman. Galton did statistical analysis of heredity 🡪 can certain traits like intelligence be inherited? He called this study eugenics and found indeed prove for the fact that certain traits are heritable.

Spearman thought that all intellective functioning is underpinned by one general mental ability called g and by specific abilities for different tasks called s. He did tests with children and thought that one single factor, g, should be responsible for all performance ratings on different tasks. And that the correlations between these tasks should be perfect. And he indeed found evidence for the presence of a general mental ability.

You can apply factor analysis on subsets or items from a questionnaire. Why would you do factor analysis?

* To find common factors, common causes. How do tests/items cluster?
* Measurement development. Like reliability, validity, diagnostic, monitoring and testing
* To correct for measurement errors
* To estimate factors scores, composite scores.



F 🡪 latent variable, unobserved

Ε 🡪 the error term, unobserved

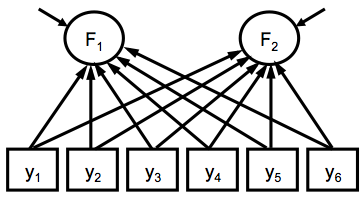
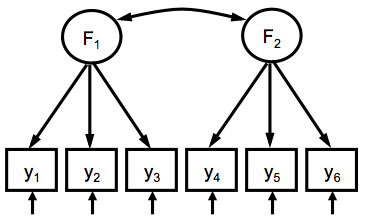
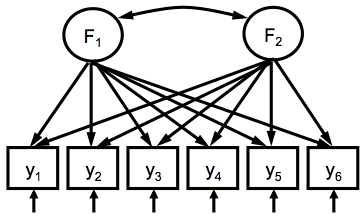
Y 🡪 the manifest variable, observed

Λ 🡪 the factor loading

θε🡪 the variance of the error term

Variables have a multivariate normal distribution. There are three factor analytical procedures:

* Exploratory factor analysis, all items load on all factors. You look whether you have enough factors. Can be tested with statistical tests. Y variables have to be correlated
* Confirmatory factor analysis, you decide which items load on which factors. You look whether you have the right factors. Can be tested with statistical tests
* Principal component analysis. There is no statistical test available. Y variables don’t have to be correlated



CFA 🡪 estimating of parameters by minimization of the residuals using ML. Check how many degrees of freedom you have. Even if the df is bigger than 0 you cannot identify the factor model without special restrictions. Because there are many possibilities and values for λ and ε to get the same df.

So to find out what the best solution is, you have to make restrictions. There are two of them:

* UVI, fix the variance of the latent variable to 1. The scale of the latent variable have become z-scores
* ULI, fix a factor loading to 1. The scale of the latent variable have become equal to the item’s scale.

Lavaan will do this for you.

Measurement validity 🡪 tests of construct validity. Goals of factor analysis is to find one or more common factors. But in order to give meaning to the factors it is necessary to establish a network, a scientific theory. The process of giving meaning to factors is called validation.

There are different forms of validity that can be tested when you have a theory and data is available. these are predictive validity, concurrent validity, criterion validity, convergent validity and discriminant validity. But two forms of validity cannot be tested:

* Content validity. Depends on the extent to which an empirical measurement reflects a specific domain of content. So you need to test the whole content, not just one part of it
* Face validity. Logical link between items and purpose, makes sense on the surface

The fit of the model is sensitive to:

* Sample size
* Number of variables in the model
* Non-normality
* Model characteristics

In order to protect against these flaws, different goodness-of-fit measures have been developed. There are two types of GOF:

* Absolute fit. Indices are based on how well the model fits in comparison to no model at all. These are chi-square (should be significant), RMSEA (parsimony adjusted, should be below 0.05), GFI (should be higher than 0.90), AGFI (same as GFI, adjusted for parsimony), RMR (if 0, perfect fit) and SRMR (should be below 0.05).
* Incremental fit. Indices do not use the chisquare per se but compare the chisquare value to a baseline model. These are NFI, NNFI, CFI (should be higher than 0.95).

To check for the GOF you specify fit.measures = T in the lavaan function. Then you get 4 types of fit. You can also use fitMeasures(fitobject) to get all indices for the GOF. How do you report GOF 🡪 the opinions differ. But you can report the ones that are in the default.

Modification indices 🡪 identify local misspecifications 🡪 modindices = T

**Lecture 3 Testing in SEM Revisited and Troubleshooting**

With SEM you can test the model as a whole and if this turns out to be significant you can also look at specific hypotheses. But only if the model is significant! So first try to find out if this is the case:

* Df = 0 and p = 1 🡪 model is not accepted or rejected, you can test the hypotheses
* Df >0 and p > 0.05 🡪 model is accepted, you can test the hypotheses
* Df > 0 and p < 0.05 🡪 model is rejected, you need to look for modification indices that are bigger than 3.84.

The chisquare test is flawed. For example, if you have a very large sample, everything can become significant. Chi2 is sensitive to:

* Sample size
* Model complexity. If you have more parameters this can lead to more type 1 errors
* Deviations from multivariate normality

Many fit indices are based on Chi2. However, some fit indices correct for sensitivity for sample size, others for sensitivity to model complexity, others for both or for sensitivity to non-normality. But what to do when fit indices differ in whether they day the model has a good fit?

The goal of testing is to find a model that is more or less correct 🡪 a model is always a simplification of reality. In order to achieve this goal, the test should warn us when the model has large misspecifications. And should not warn us when a model has small misspecifications. But is it possible to detect large misspecifications and ignore small misspecifications? You can answer this question with a Monte Carlo simulation.

In this simulation you have a population model (a correct model) on which you create a data set. And then you analyze this data with incorrect models. Which means that you should get an error! These models differ in the size of the correlated error and the size of the factor loadings. Larger misspecifications should correspond to larger Chi2 and RMSEA, and a lower CFI. The same size misspecification should lead to the same size of Chi2, RMSEA and CFI.

But is this what really happens? Larger misspecifications, errors, correspond to a larger Chi2, RMSEA and a smaller CFI. But the same size misspecification does not correspond to the same size Chi2, RMSEA and CFI. So the answer is 🡪 yes, it’s possible to detect large misspecifications and ignore small misspecifications. But the opposite is also possible! Detect small misspecifications and ignore large ones. So there is a relationship between the factor loadings and whether your model is accepted or rejected 🡪 but It’s not a one-on-one relationship.

The power of a test depends on:

* Sample size
* Complexity of the model
* Deviations from non-normality

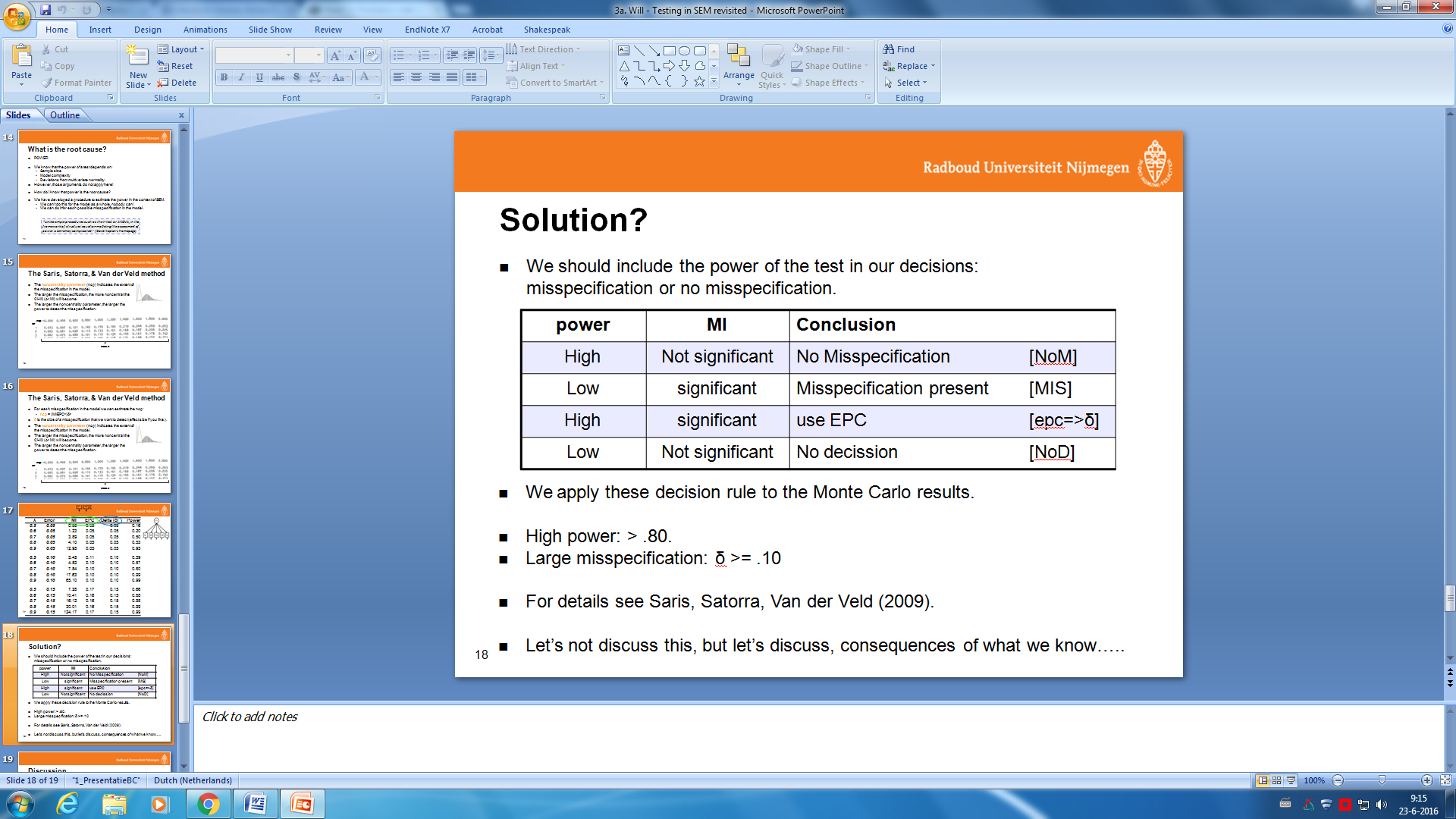
However these are not the arguments that apply here. Power is the root cause 🡪 you can look at the power for each possible misspecification. You cannot do this for the model as a whole.

For each misspecification you can estimate the noncentraility parameter, ncp. This ncp indicates the extent of the misspecification in the model 🡪 the larger the ncp, the larger the misspecification and the more noncentral the Chi2 will become. The larger the noncentrality parameter, the larger the power to detect the misspecification.

Ncp = (MI / EPC2) δ2

δ is the size of a misspecification that we want to detect 🡪 effect size.

You should include the power of the test in you decisions 🡪 misspecification or no misspecification.



Goodness of fit statistics don’t perform as expected. So you should know the conditions that affect the model fit:

* Deviations from normality
* Model complexity
* Sample size
* Size of unrelated model parameters, in this case high factor loadings

Good models can be rejected and bad models can be accepted if any one of these conditions is at work!

But should we therefore stop evaluating model fit? No just know what the sources are that can affect your fit statistics. Learn whether any of those source is present in your data and assess how much your results are affected.

Trouble shooting 🡪 what to do when data don’t behave?

Practical issues in SEM:

* Sample size 🡪 bigger is generally better, it means you get more reliable estimates
* Missing values 🡪 look at patterns, mechanisms and robust algorithms.

Basic assumptions of SEM:

* Multivariate normality, no outliers
* Absence of multicollinearity and singularity (extreme high correlations)
* Normality of residual covariances, residual covariances should be small and symmetric

You need to screen for outliers and missing values. Check the univariate distributions and bivariate associations. So you should look at skeweness, kurtosis and singularity. So you should try to test these multivariate assumptions before you conduct your analysis.

What to do with outliers? You can either ignore them or delete them. Another options is to winsorize the extreme values 🡪 this is a procedure that draws extreme scores closer to the mean. It’s the transformation of statistics by limiting extreme values in the statistical data to reduce the effect of possibly spurious outliers. So you keep the cases, but they are modified. So you set a boundary, for example +/- 3 SD and then you winsorize all the values above that boundary.

Winsoration uses percentiles to detect outliers. A 90% winsoration identifies all the data below the 5th percentile and set it to the 5th percentile and it sets data above the 95th percentile to the 95th percentile.

Winsoration can be a viable option, especially with smaller samples (smaller than 500) and only a couple of outliers. However, rank ordering among the outliers is lost in winsorizing 🡪 so with many outliers and large samples this can be a problem. Proactive data collection strategies can reduce issues with extreme values 🡪 look at participants that have questionable data.

When is missingness a problem? The general rule is that when you have 5% or more missing cases you should investigate it. Patterns of missingness make a differences! For example 20% of missing cells can be 75% missing cases.

There are different levels of missingness:

* Items, leaving an item or two blank. Assuming you want to aggregate missing items can be difficult. You should compute the mean with all available items 🡪 do not compute the sum! But make sure you still represent the right scale.
* Scale, omitting answers for an entire scale or construct. If it’s not MCAR, you should use ME or MI to prevent biased results. Differences in missing patterns can be tested
* Survey, failing to complete an entire survey. There’s not a lot you can do about this, look whether you have a selection bias by using systematic non-response parameters, SNP’s.

Proactive 🡪 missingness can be prevented, to some degree. You can give advanced warnings, personalized surveys, follow-up reminders and monetary incentives. Reactive 🡪 using analytical strategies to control for missingness. Like unwise deletion (listwise and pairwise), single imputation (mean substitution, LOCF) or multiple imputation (ML and MI).

In repeated measures you can look at patterns of missingness over time. Tehse can e:

* Monotone. Dropout 🡪 when data are available at every assessment until a time when the participant drops out and provides no further assessments
* Intermittent. When there is a missing observation in between observed assessments
* Mixed. Both monotone and intermittent missingness

The same mechanisms, MCAR, MAR, NMAR, apply to missingness in studies involving repeated measures 🡪 with the same consequences regarding bias and power.

Attrition analysis 🡪 you can look at who are more likely to drop out, attrition. Most methods have been developed to identify the absence of MCAR in data with monotone missingness. Does attrition occur randomly? This can be inferred from various logistic regression techniques. Attrition tends to be MAR. Which makes ML and MI extremely advantageous in repeated measures.

Missingness can lead to regression toward the mean if response rates increase.

Assumptions of normality:

* Independent observations
* Large sample size
* Correctly specified model
* Multivariate normality
* Continuous data

If all these conditions are satisfied then normal maximum likelihood estimates are accurate, efficient and consistent.

In reality data are rarely normally distributed. Minor deviations are probably okay. But severe deviations can lead to biased results. You can test univariate normality with skewness, kurtosis and shpiro- Wilkes test. You can test bivariate normality with scatterplot, q-q plot and perspective and contour plot. You can test multivariate normality with Mardia’s, Henze-Zirkler or Royston’s. But you can only do this with complete cases. To test these different types of normality you can use the MVN package.

Transformations can be handy except when:

* You don’t have a transformation that normalizes the distribution
* Measures in a single analysis require different types of transformations
* Interpretation of the unstandardized coefficients, M and SD, is important.

Other options are:

* Weighted estimators. ADF is asymptotic distribution free 🡪 aka weighted least squares, WLS. Residuals are weighted based on asymptotic covariances and this should give you less biased estimates. But requires very large samples! And it tends not to converge for complex models. Rarely usefull
* Robust estimators. These decrease the computational intensity of the ADF estimator. Huge samples are not necessary. Useful in many situations
* Bootstrapping. Iterations, you sample with replacement. So you use slightly deviations from the same sample multiple times. Estimates from bootstrapped samples form a distribution of statistic of interest. SE’s calculated from bootstrapped estimates. Is especially useful with smaller samples and indirect effects

For Satorra-Bentler scaled chi2 and SE’s you need complete data. This procedure uses ML estimates, so avoids issues with WLS. Uses sandwich estimator to adjust the chi2 and the SE by a factor based on the degree of non-normality. With Yuan-Bentler you can use complete and incomplete data. Is a robust estimator. Also uses ML estimates, so avoids issues with WLS. Uses different sandwich estimators to adjust the chi2 and SE by a factor based on the degree of non-normality.

When should you use bootstrapping?

* When the theoretical distribution of a statistic of interest is complicated or unknown 🡪 like indirect, mediation, effects.
* When the sample size is insufficient for straightforward statistical inference.
* When power calculations have to be performed and a small pilot sample is available

**Lecture 4 Multiple Groups Analysis**

Interactions are used to test moderation. First you have to center all the continuous measures and then you have to create the interaction term yourself. For categorical predictors you need to use proper coding schemes. In lavaan the interactions need to be created and added to the data beforehand 🡪 you cannot multiply predictors in the model like in lm().

If a categorical variable represents independent groups then SEM can be estimated separately for each group. Differences between parameters can be tested to determine if groups differ. This is done with equality constraints 🡪 in which you force regression coefficient to be the same for different groups.

But how do you do this? First you have to tell lavaan that you have different groups by typing groups = ‘variable’. It’s the same as splitting afile in SPSS. Then the constraint forces the regression coefficients to be identical for each group 🡪 the model has now 1 df. Specify c(a1, a1) \* V1 + V2 in your model. You can also specify multiple equality constraints in which you say group.equal = c(‘regressions’). Then you can compare groups of parameters.

What can be constrained?

* Loadings 🡪 factor loadings on the observed variables
* Intercepts 🡪 intercepts of the observed variables
* Residuals 🡪 residual variances of the observed variables
* Means
* Covariances

You can just define this in lavaan

Measurement invariance 🡪 insuring that the latent variables mean the same for your groups. The model should mean the same thing for the groups that you have. So with measurement invariance you look at different things. Four models should be estimated:

* Configural invariance 🡪 does the model fit the data separately for each group? Had no constraints. Groups = ‘variable’
* Metric invariance, weak 🡪 are the factor loadings the same for each group? Factor loadings are constrained. Group.equal = c(‘loadings’)
* Scalar invariance, strong 🡪 are the factor loadings and intercepts the same for each group? Factor loadings and intercepts are constrained. Group.equal = c(‘loadings’, ‘intercepts’)
* Strict invariance 🡪 are the factor loadings, the intercepts and the residual variances the same for each group? Factor loadings, intercept and residual variance are constrained. Group.equal = c(‘loadings’, ‘intercepts’, ‘residuals’)

Models are nested so can be tested with chisquare.

With measurement.Invariance() lavaan will automatically test these models. This function comes from the semTools package. You have to specify the model, the data and the group variable.

Testing invariance is important! You should not make some group comparisons. But testing invariance can be complicated 🡪 there are differences in terms and recommendations. There is also partial invariance. Testing invariance with lavaan is easy, but you need to understand MI.

There are several steps in scale development:

* Selection of the ‘best’ items on face validity.
* Assessment of psychometric properties 🡪 perform factor analysis (validation of the dimensional structure). And assess the scale-reliability, e.g. with Cronbach’s alpha. Plus perform (construct) validity tests.
* Measurement equivalence 🡪 is property of a measurement instrument. If you can do comparisons across time and groups. So if a certain scale works equally well for different groups.

You look at the invariance of the measurement properties 🡪 like factor structure, validity, reliability. If it is indeed invariant then you cannot make comparisons! Reponses on the same question should be comparable between persons and over time. You need a test that looks whether an instrument is equivalent over groups.

Such a test should reveal whether people in different groups:

* Express themselves on the same dimension, length
* Use the same metric
* Use the same reference point

This is studied in multi-group CFA. If you have measurement equivalence what does this mean?

* If configural invariance is correct then the same thing is measured in all groups 🡪 length. Both scales measure the same concept in each group. But this does not mean you can compare the, it might be biased
* If metric invariance is correct, the scale-point distance is equal across groups. It reflects the same metric. You can compare the relationships between variables across groups
* If scalar invariance (which adds the mean) is correct the zero-point is on the same place in all groups 🡪the means of the variables can be compared across groups. The item intercepts are the same.

**Lecture 5 Causal Modeling and Mediation**

The goal of causal analysis is to correctly estimate causal effects. Under what conditions will b1 be estimated correctly? And can we interpret b1 as the causal effect?

x1

y1

β1

ζ1

Important aspects in causal analysis are:

* The disturbance term
* A spurious effect
* Equivalent groups

The disturbance term, contains all causes of Y1 🡪 the unexplained variance. If you add a variable X1 to the model, then the disturbance term contains all causes of Y1 except X1. If the variables are standardized the disturbance term equals 1 – R2.

A spurious effect is an effect that changes when you add a variable to the model. so the effect of X1 on Y1 changes if you add X2. Spurious effects can be the result of:

* Suppressor variables, when predictors are correlated
* Confounding variables, that have common causes
* Intervening variables, mediators

Group equivalent occurs when groups are the same. The distribution of the variables are the same. For example when you have an equal amount of males and females in each group. Then if you manipulate something in group A 🡪 any difference between the groups must be the result of the manipulation. Because both groups were the same before the manipulation!

There are many definitions in causality. But causality contains three important components:

* Isolation
* Association between X1 and Y1
* Direction of force from X1 to Y1

An association is not enough 🡪 correlation is not causation

If there is association and isolation 🡪 correlation is proximate to causation

But if you also have direction of force 🡪 correlation is causation.

Isolation by turning of all other causes is an unobtainable ideal 🡪 that’s why research often uses pseudo-isolation 🡪 the disturbance term is not correlated with X1. When you violate pseudo-isolation, then B1 will be incorrectly estimated. Because then the disturbance term contains a cause X2 that is correlated with X1. A solution to this issue is to use causal research designs.

Causal research designs:

* Experimental designs. In this design you manipulate the cause and you use random assignment
* Quasi experimental designs. In this design you manipulate the cause but you don’t use random assignment. Instead you use a control group
* Non-experimental designs. In this design you don’t manipulate the cause and you don’t use random assignment. But you use a control group.

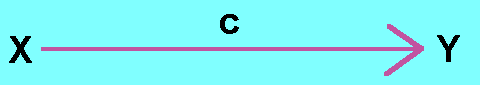
Manipulation is necessary to establish direction of force. And randomization or control help to approximate pseudo-isolation. Random assignment evenly distributes the effect of all other causes, the disturbance term, over the treatments. So no correlation between X1 and the disturbance term! This is how you get pseudo-isolation.

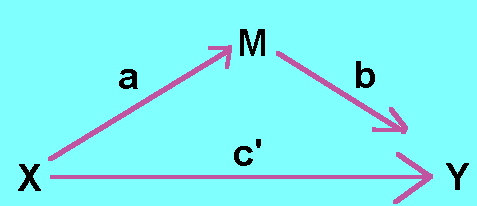
With SEM you can analyze experimental and quasi experimental data. And even non-experimental data! You have to control for variables that can violate pseudo-correlation. You add variables to the model that will cause a spurious effect 🡪 finding these three types of variables is called causal modelling. You can use the literature and theoretical knowledge to try to find these variables.

So normally you use manipulation to control for direction of force, but in non-experimental designs you use theoretical knowledge.

Causal theories can be rejected! You can claim that there are multiple causal models to explain the data. This means that two models can be equally good in explaining the data.

Mediation 🡪 there’s a relationship between X and Y and there’s another variable or variables that explain this relationship. They explain the underlying mechanisms of the relationship. But how do you test mediation?

 c 🡪 total effect

 c’ 🡪 direct effect

Path AB is the indirect or mediation effect and is defined as the reduction of the effect of X on (c – c’). So AB = c – c’ when you use the same participants in each regression equation. Is mediation ever complete?

* Complete mediation. X no longer affects Y at all. (c = 0) 🡪 rarely occurs in regression models
* Partial mediation. The effect of X on Y is still significant but is reduced. It is however still different from 0

What is the difference between mediation and an indirect effect? Testing mediation without a significant effect between X and Y is debatable and not recommended. Then you can only test an indirect effect 🡪 not mediation. There are two ways to test mediation.

Old school 🡪 doing 4 separate regressions:

* Step 1 is to show that X predicts Y. Test the total effect, path c. This establishes that there is an effect that may be mediated
* Step 2 is to show that X predicts M. Testing path A
* Step 3 is to show that M predicts Y. Testing path B
* Step 4 is to show that M reduces the effect of X on Y, to zero. Testing the direct effect, c’. Total effect minus the direct effect = indirect effect (AB).

Do all of these steps have to be met? It helps is step 1 is met, but it is debatable if it’s required. Step 2 and 3 are essential in establishing mediation. Step 4 does not have to be met, unless you are testing complete mediation.

Problems with the causal steps approach:

* It suggests causality!
* It misses some true effects, there are biased results and a low power often occurs
* It is not based on a quantification of th every thing it is attempting to test! It does not test a\*b pathways.

With the Sobel test you can determine the significance of ab. But the sampling distribution of the ab pathway tends to be kurtosed and skewed 🡪 Sobel test is very conservative. So there are problems with this test.

New school 🡪 bootstrapping by adjusting the SE’s of the coefficients provides more reliable results. Is more valid and more powerful 🡪 based on the indirect effect itself. It makes no assumptions of the sampling distribution of the ab pathway. So use bootstrapping! Some general mediation design issues:

* Mediation should be close in time to X or Y
* There can be alternative models. Y predicts X? M predicts X? Y predicts M?
* There can be measurement errors 🡪 latent variables have advantages
* Omitted 3rd variables
* Moderated mediation or mediated moderation

Causality 🡪 mediation is used across many psychological areas. You need parameter labels, indirect effects need to be defined. Also you need to look at the total effects which equals the direct and indirect effect. To check this add := c + (a \*b).

Multiple mediation 🡪 make sure you multiply the correct pathways

Mediation is a powerful tool. But mediation should be assessed after the predictor and before the outcome, otherwise it’s all correlation! Use bootstrapping.

**Lecture 6 Longitudinal Data Analyses**

Lord’s Paradox 🡪 a paradox in the interpretation of group comparisons. You have different modeling techniques to asses changes over time. But these techniques provide different results. One study investigated whether students gain weight during their first year of college. There are two ways of analyzing the data, which yield different results. According to Pearl statistician 1 examined the total effect of gender on weight gain. But statistician 2 examined the direct effect of gender on weight gain, adjusting for initial weight. So controlling for whether the two groups differed at T1.

Investigating residual change:

* Variable-oriented method. You assume the population is a homogeneous set which all fall along a homogeneous distribution. Then you use autoregressive models, Markov models. Cross-lagged models.
* Person-oriented method. You assume the population is a heterogeneous set which have different distributions. Then you can use latent growth mixture models, trajectories. Longitudinal cluster analysis.

Autoregressive models 🡪 each variable at t(n) is a function of a variable at t(n-1) and nt of any variable before. This is called a Markov process. Coefficients indicate the degree of stability of individual differences. How stable is this variable from T1 to T2 and from T2 to T3? But you can extend this model! With cross-lagged autoregressive models you also make cross-paths of each variable predicting changes in the other variable. You can look at bi-directionality. Does a change at T1 to T2 predict X1 or Y1? Cross paths can reveal other effects, like delayed effects of certain variables.

However these models have some modeling issues:

* Model evaluation 🡪 correlating error terms. You cannot explain variance at T1 and explain the same variance at T2. These unexplained variances are correlated! The model parameters reflect shared unexplained variances 🡪 which will improve the fit of the model.
* Measurement invariance 🡪 invariance also applies to time. Testing is less automated in lavaan. Measurement invariance isn’t just for testing group differences. Loadings, intercepts, regressions and variances can also be invariant across time! To test this you will need to constrain different paths to be equal 🡪 instead of constraining the same path to be equal across different groups. You fix the unstandardized regression coefficients to be the same from T1 to T2.

If your model does not fit you can try another modelling technique or add more paths. Less stable variables will less likely predict anything else. In these variables there’s more variance left to be explained.

LGCM 🡪 latent growth curve model.

LCGM 🡪 latent class growth model. Person-oriented method

LGCM’s can be performed using mixed effects models or with SEM. You can capture random effects by latent variables. LGCMS require at least three time points for each individual. And you fit a straight line for each person 🡪 so you test a linear model. If you want to test a curve linear model you need 4 time points for each individual.

Relationship to multilevel modeling 🡪 equivalent if the ML option is chose. Advantages of SEM:

* Measures of absolute fit
* Easier to respecify, there are more options for respecification
* More flexibility in the error covariance structure
* Easier to specify changes in slope loadings over time
* Allows latent covariates
* Allows missing data in covariates

Advantages of MLM:

* Better with time-unstructured data
* Easier with many time points
* Better with fewer participants
* Easier with time varying covariates
* Random effects of time varying covariates allowable

Slope is the rate of change and some people are changing more than others and have larger slopes. Some people are improving or growing. Some are declining and some are not changing. Intercept is where the person starts, people can start at different initial levels. The error is then how far the score is from the line.

For both the slope and intercept here is a mean and a variance. But how do you interpret them in LGCM?

* Mean 🡪 for the intercept this means where does the average person start? And for the slope it means what the average rate of change is
* Variance 🡪 for the intercept this means how much individuals differ in where they start. And for the slope it means how much individuals differ in their rage of change 🡪 different slopes for different folks.

Modeling issues with LGCM:

* You should not use CFI and TLI, unless the independence model is recomputed. You cannot trust the goodness of fit
* Never have the intercept cause the slope factor or vice versa. You never use an intercept to predict a slope of the same variable
* Do not interpret standardized estimates except the slope-intercept correlation. The SE are not to be trusted.

Quadratic slope factor loadings are the square of linear slope factor loadings. You can also have logarithmic slope for factor loadings 🡪 you would just take the logarithmic of the linear slope factor loadings.

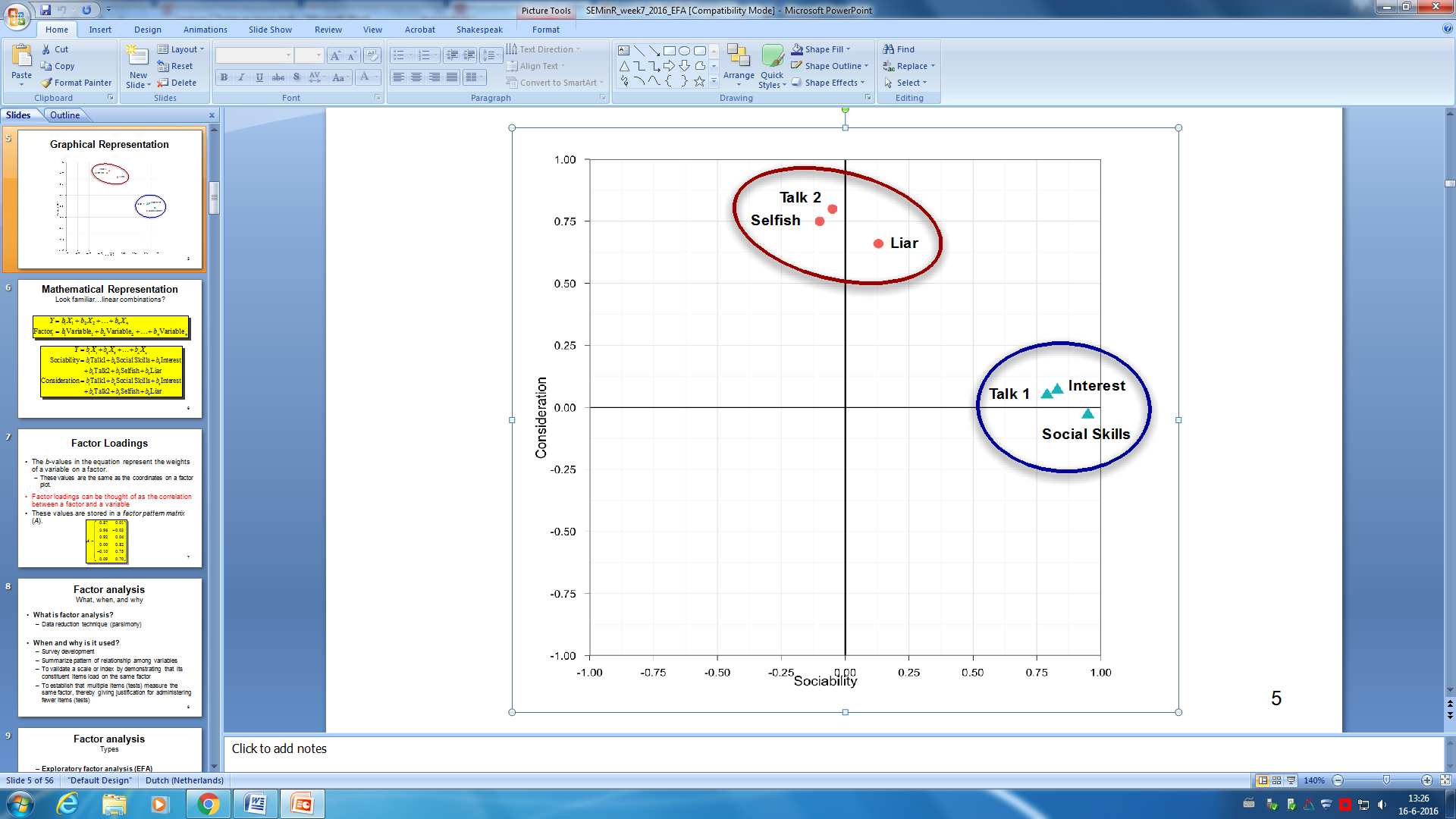
Variances give information about whether people differ in their slopes 🡪 individual variation in starting points or on certain variables.

Different models can give different results, have different interpretations. Models can be further investigated by looking at distributions and whether you should transform your variables. You can test constraints or recalculate GOF indices. And maybe simplification or extension of your model also works.

**Lecture 7 Tying up loose ends. EFA, PCA and internal reliability**

Factor 🡪 if several variables correlate highly, they might measure aspects of a common underlying dimension. These dimensions are called factors, constructs or latent variables. But the components are different from factors!

Factors are classification axes along which measures can be plotted. The greater the loading of variables on a factor, the more that factor explains relationships between those variables. You can represent the items graphical and then you can see in which factors the items are falling:



The factor loadings are the correlation between a factor and a variable/item. These values are stored in a factor pattern matrix. The b-values in the equation represent the weight of a variable on a factor, these are the same as the coordinates in a factor plot.

Factor analysis is a data reduction technique. You reduce the data to a smaller number of factors. You often use it in:

* Survey development
* Summarizing patterns among variables
* Scale validation, showing that different items load on the same factor
* Establishing that multiple items measure the same factor. So you can use fewer items.

There are two types of factor analysis 🡪 EFA and CFA. In EFA you don’t have a theory yet, so you build one. It commonly uses an extraction technique known as principal component analysis, PCA. In CFA you already have a theory which you are testing. It requires a SEM framework, lavaan.

Common variance is the variance that a variable shares with other variables. Unique variance on the other hand is variance that is unique to a particular variable. The proportion of common variance in a variable is called communality:

* Communality = 1 🡪 all variance is shared
* Communality = 0 🡪 no variance is shared
* 0 < communality <1 = some variance is shared

PCA and FA calculate this communality differently. PCA assumes all variance is shared 🡪 so communality = 1. This is a formative measurement. It assumes that items do not necessarily correlate, the items form the latent construct. FA estimates the communality, uses squared multiple correlation. It assumes that items are correlated 🡪 items reflect the latent construct. This is a reflective measurement.

Steps in EFA (PCA)

* Preliminary analyses 🡪 identify sampling adequacy. Examine the correlation matrix.
* Extract/determine the number of factors
* Rotate the factors and interpret the results.

GI 🡪 GO 🡪 the quality of analysis depends upon the quality of the data. It depends on the number of items and factors 🡪 you need a large sample size. Know what items you put in the analysis! Variables should be positively correlated, but avoid multicollineairty and singularity!

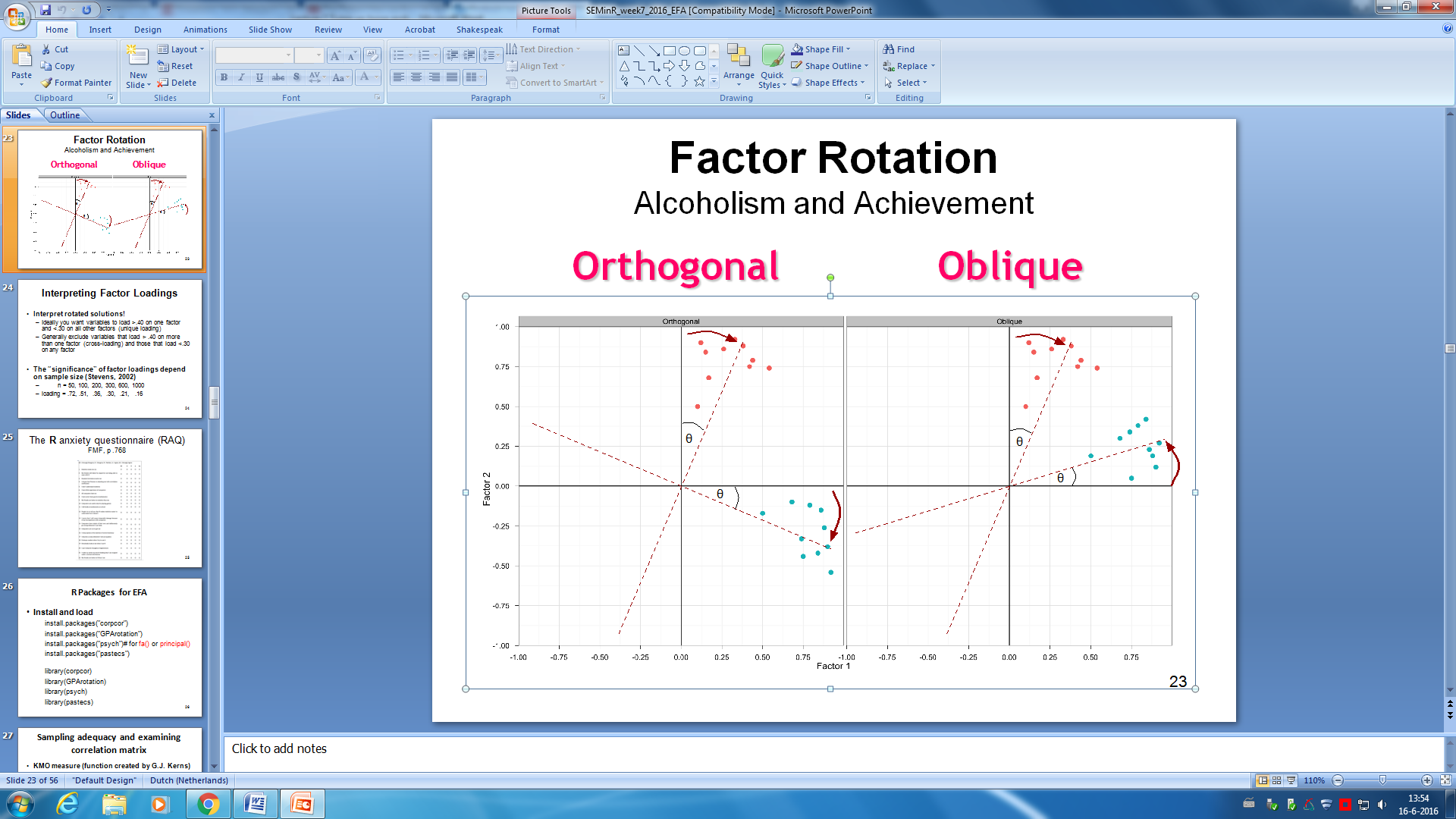
* If all correlations are above .3 then consider using FA
* If all correlations are below .3 then consider using PCA

Kaiser-Meyer-Olkin measures sample adequacy, is the squared correlation/squared partial correlation. Should be greater than 0.5. Bartlett’s test tests sphericity 🡪 whether variables in the matrix are correlated, should be significant at p < .05. The determinant is an indicator of multicollinearity, should be greater than 0.0001.

Extraction 🡪 is the process by which factors are determined form a larger set of variables. PCA is often used for EFA and ML is often used for CFA. The goal is to extract factors that explain as much variance as possible with as few factors as possible. Determining the number of factors:

* Kaiser criterion. Look at ss loadings that are bigger than 1, this is the Kaiser criterion. So you keep factors that have eigen values, factor loadings, bigger than 1.
* Scree plot. The scree plot gives a visual representation of the variance explained. You can use the point of inflexion of the scree plot to determine how many factors you should keep. This inflexion is the elbow in the curve.
* Communalities. Can help you determine when to stop adding factors.

Factor rotation 🡪 makes interpretation easier. You use orthogonal rotation when factors are uncorrelated, varimax is preferred over quartimax. You use oblique rotation when factors intercorrelate, promax is preferred over oblimin.



Ideally you want variables to load more than .40 on one factor and less than .30 on all other factors. This represents a unique loading. Generally you exclude variables that load more than .40 on more than one factor, this is called a cross loading, and variables that load less than .30 on any factor. The significance of factor loadings depends on sample size.

Nfactors is the same number as the amount of items. With principal() you can create a model and then you can extract the factors and run it again. Once you identify the amount of factors you can rerun the analysis, the model and specify, and specify this amount.

Factor residuals 🡪 difference between the reproduced matrix and the actual correlation matrix.

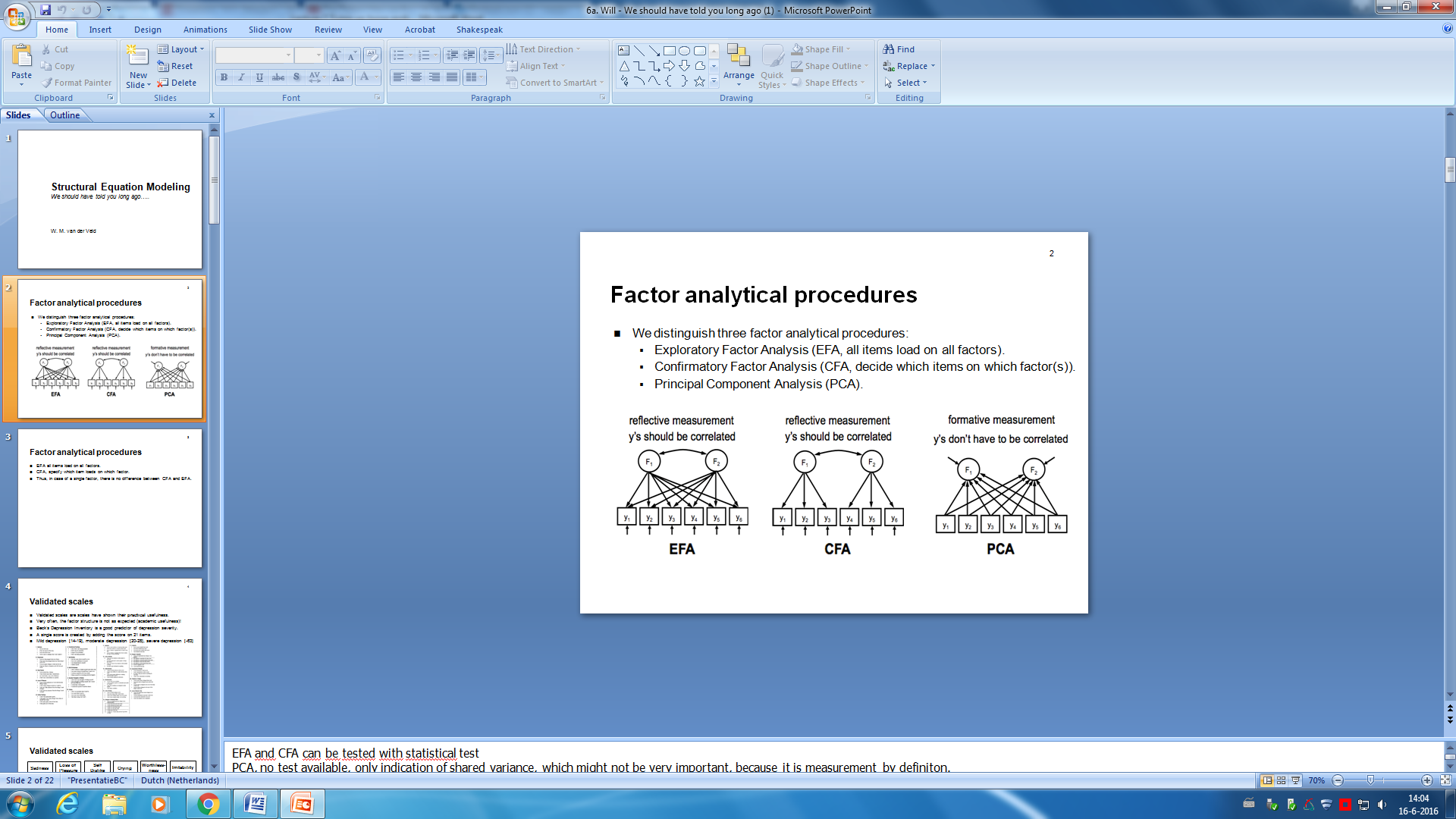
Once you represented the residuals as a single column of data you can calculate the number and percentages of large residuals. Also look at the mean, this should be small. Examine the residual distribution with histograms. If there are no large residuals you can rotate 🡪 specify the rotate option in the principal() function. If not specified, varimax is the default option. This is complex to interpret so you can use print.psych() to sort the items.

Negative loadings mean a negative correlation, often you need to reverse the coding of the variable. Don’t remove multiple items at once! Remove 1 items and then rerun the model to see if it changed. Re-run until you reach a clean solution.

Reliability 🡪 cronbach’s alpha is not the only way to look at reliability. There are alternative packages like omega.

There are three factor analytical procedures:

* EFA, all items load on all factors
* CFA, you decide which items load on which factors. Thus in case of a single factor, there is no difference between CFA and EFA
* PCA



Validated scales 🡪 already showed practical usefulness. But often the factor structure is not as expected 🡪 the academic usefulness is not as good. This is common for many scales, that things come out differently and that the structure can differ. This does not mean that you cannot use the scale. But what is the structure that you should use? What is the correct structure?

Equivalent models 🡪 just by looking at the item content you can come up with different structures. Statistics will not help you to decide between the different structures. Several can be useful. You need a theoretical framework to decide if the fit of the different structures is the same. All models could have equally produced the observed data 🡪 these models are called equivalent models.

There are always, theoretically, alternative explanations for the same data. This is problematic!

The promise of SEM is that we can correct for measurement error by analyzing factor models. Which is good 🡪 because measurement errors can result in inconsistent and biased estimates! However in reality you can have many practical obstacles to correction for measurement error like:

* Sample size. You need many participants. Otherwise easy rejection
* Complexity. If you have many variables your model is often rejected

You can for example make a model easier by looking at composite scores. Then you can have a smaller sample size. You use factor estimation and adding up or removing items to strive for a simpler structure. Factor scores are better estimated when the simple structure holds. However the sum scores of the factor scores are also not free from measurement error.