# LATENT VARIABLE MODELS

Session 2: Basic CFA models

## Session 2 - Basic CFA Models

Contents this afternoon:

- Reflective and formative factor models
- Model estimation, evaluation, modification
  - Parameter estimation
  - Assessing model fit
  - Model modification

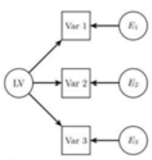
## Latent variables

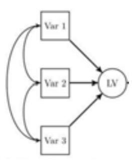
- Latent variables (<u>LV</u>s) are variables that are not directly observed, but are inferred from other variables that are directly observed (<u>OV</u>s)
- LVs represent a construct or concept that researchers are interested in, but cannot directly measure:
  - E.g., depression, anxiety, aggressiveness, socio-economic status, wellbeing, quality of life, social skills, intelligence, mathematical abilities, …
  - In this workshop: focus on continuous LVs
    - LVs can also be categorical (latent classes), but outside scope of this course

## Factor analysis

- Confirmatory factor analysis (CFA)
  - We have a (relatively) clear idea about:
    - number of factors underlying observed variables
    - with which observed variables they are related
    - what they represent
- Exploratory factor analysis (EFA)
  - When we have no clear idea about that
  - Not in this course
- Both assume arrows to go from factor to indicator (i.e., reflective model)

## Reflective and formative LVs




- Reflective LV:
  - OVs reflect the LV
  - LV ('underlying' factor) 'causes' the OVs
  - LV is exogenous
  - OVs are endogenous
  - OVs must be correlated
  - e.g., depression, intelligence, …
- Formative LV:
  - OVs form the LV
  - OVs 'cause' the LV
  - LV is endogenous
  - OVs are exogenous
  - OVs not necessarily correlated
  - E.g, SES, physical health (balanced diet, regular exercise, sufficient sleep)

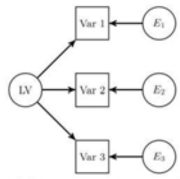
Errors are also LVs, but do not represent a construct, and are always exogenous

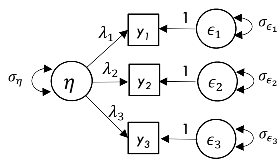
## <u>Reflective</u> measurement model

The score on item *i* of person *j* is given by:

$$y_{ij} = \tau_i + \lambda_i \eta_j + \epsilon_{ij}$$

Often, we assume a centered model (i.e., all means are zero), so $\tau_i$ (item intercepts) are zero and can be omitted:



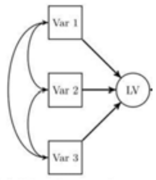



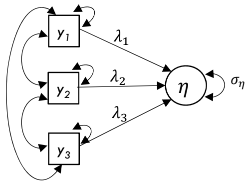
## Formative measurement model

The score on the latent factor of person $j$ is given by:

$$\eta_j = \tau + \lambda_i \, y_{ij}$$

Often, we assume a centered model (i.e., all means are zero), so $\tau$ (intercepts) is zero and can be omitted:



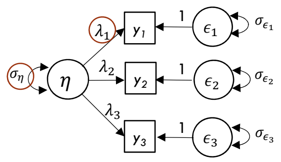

## Coefficients

- A factor loading is a regression coefficient:
  - **Unstandardized** factor loading:
    - expected increase in OV, when LV increases by 1 (reflective model)
    - Expected increase in LV, when OV increase by 1 (formative model)
  - **Standardized** factor loading:
    - bivariate correlation between OV and LV
    - expected increase in SDs of OV, when LV increases by 1 SD (reflective model)

## Identification

- E.g., we have 1 reflective LV, with 3 indicator variables
- There are 6 pieces of information about the scales and associations of the variables in the sample data
- In the (population) model, there are 7 unknowns (parameters) to estimate
  - We assign a constant value to ('fix') one of the parameters
  - Then values of other 6 parameters can be freely estimated, using the sample information
    - In other words: this yields a unique solution -> the model is identified

| | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $X_1$ | Var($X_1$) | | |
| $X_2$ | Cov($X_1$, $X_2$) | Var($X_2$) | |
| $X_3$ | Cov($X_1$, $X_3$) | Cov($X_2$, $X_3$) | Var($X_3$) |




## Identification: Reflective LVs

- Minimum requirements for identification of reflective LVs – rules of thumb:
  - > 3 indicator variables per LV (preferred)
    - Scale of LV has to be set by fixing a single parameter
    - Some errors are allowed to correlate
  - 3 indicator variables per LV
    - Scale of LV has to be set by fixing a parameter
    - No error covariances
  - 2 indicator variables per LV
    - Scale of LV has to be set by fixing a parameter
    - No error covariances
    - Both loadings set to equality
  - 1 indicator variable per LV
    - Better use observed variable, without underlying LV

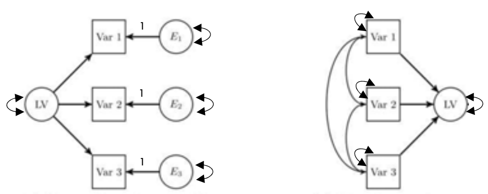
## Identification: Reflective LVs

3 ways to identify scale of an LV:

1. Standardize LV: fix LV's variance to 1
   - In lavaan: use model syntax, or set 'std.lv = TRUE' in cfa() function
2. Marker variable: set factor loading of an item to 1
   - Best practice: use the item most strongly correlated with the factor
   - Most common practice: use first item (not a major sin but always check if marker item is substantially correlated with factor)
     - Default in lavaan's cfa() function
3. Effects coding: set sum of loadings equal to the number of indicator variables
   - See example 3.3.1 in Beaujean book
   - Very rarely used, so skipped in this course

Yield same *standardized* solution, but different *unstandardized* solutions.

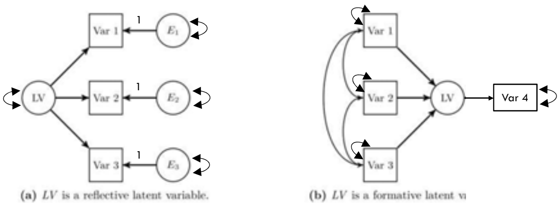
## Identification: Reflective vs. Formative LVs




(a) $LV$ is a reflective latent variable. (b) $LV$ is a formative latent va

| | |
|---|---|
| Reflective model can be identified through a single restriction: then # of parameters to be estimated = # of sample statistics | Formative model is not identified here. Formative measurement models require at least 1 additional variable, caused by the formative LV, to be identified. |

## Identification




(a) $LV$ is a reflective latent variable.

(b) $LV$ is a formative latent va

| | |
|---|---|
| Reflective model can be identified through a single restriction: If we fix a factor loading or variance to 1, there are 6 sample stats and 6 params (arrows) to be estimated | Formative model with additional variable is identifiable through a single restriction: If we fix a factor loading or variance to 1, there are 10 sample stats and 10 params (arrows) to be estimated |

## Identification

- If the model involves (co)variances only (i.e., no mean structure)
  - If # OVs in the model = $P$, then # of sample statistics = $P(P+1)/2$
  - So max. # of (model, population) parameters that can be freely estimated with $P$ observed variables is $P(P+1)/2$
- SEM models can be:
  - **Just identified**
    - No. of free parameters = $P(P+1)/2$
    - Model always fits data perfectly
  - **Underidentied**
    - No. of free parameters > $P(P+1)/2$
    - Free parameters cannot be estimated, because there is no unique solution
  - **Overidentified**
    - No. of free parameters < $P(P+1)/2$
    - All free parameters can be estimated. Generally, model fits data imperfectly -> degree of model fit can be quantified and compared between models

## Identification

Two basic conditions for model identification:

1) The number of free(ly estimated) parameters in the model ≤ the number of non-redundant (unique) elements in the sample variance-covariance matrix
2) Each latent variable needs to be assigned a scale (i.e., mean and variance)

Thus:

- In SEMs with OVs only, the model is always (just- or over-) identified
- In models with LVs, some parameter values have to be fixed to a constant by the user for the model to be identified
- Further assumptions:
  - Normality: all latent variables (latent factors and residuals/errors) are normally distributed (thus note: OVs need not be normally distributed)
  - Linearity (associations between variables in the model are linear)

## Identification

- With overidentified models, we can select a 'best' model by comparing the models' trade-offs between
  - Models' misfit to the data
    - Closer fit (less misfit) to data is better
    - Quantified by chi-square value
  - Parsimonity
    - More parsimoneous is better (Occam's razor)
    - Quantified by df

- $df$ = # knowns - # unknowns
  - # of sample stats (knowns) - # of free model parameters (unknowns)
- Just identified models have $df$ = 0
- Overidentified models have $df$ > 0
- Under identified models have $df$ < 0 (cannot be estimated)

## Question

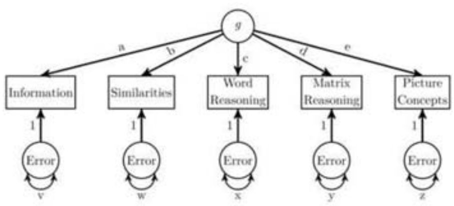



Figure 3.3 Single-factor model of five Wechsler Intelligence Scale for Children-Fourth Edition subtests.

How many sample (co)variances are there?
How many population parameters to be freely estimated?
How many degrees of freedom?

## Examples and exercises

- Example 3.3 and 3.4 – Part I

## SEM parameter matrices

- This mornings examples involved a **structural model** only:
  - **β**: a matrix of regression coefficients (single-headed arrows)
  - **Ψ**: a matrix of (co)variances not explained by the regression equations (double headed arrows)
- SEMs with LVs also involve a **measurement** model:
  - **Λ**: a matrix of factor loadings, relating observed variables to reflective latent variables
  - **Θ**: a matrix of measurement error variances

## SEM parameter matrices

- When SEMs involve both a **measurement** and **structural** model, model-implied covariance matrix is given by: $\hat{\mathbf{\Sigma}} = \mathbf{\Lambda}(\mathbf{I}-\mathbf{\beta})^{-1}\mathbf{\Psi}\left[(\mathbf{I}-\mathbf{\beta})^{-1}\right]^{\mathrm{T}}\mathbf{\Lambda}^{\mathrm{T}}+\mathbf{\Theta}$
- If model involves **measurement** model only, this simplifies to: $\hat{\mathbf{\Sigma}} = \mathbf{\Lambda}\mathbf{\Psi}\mathbf{\Lambda}^{\mathrm{T}}+\mathbf{\Theta}$
- If model involves **structural** model only, this simplifies to: $\hat{\mathbf{\Sigma}} = (\mathbf{I}-\mathbf{\beta})^{-1}\mathbf{\Psi}\left[(\mathbf{I}-\mathbf{\beta})^{-1}\right]^{\mathrm{T}}$

## SEM parameter matrices

If *P* is the number of observed variables and Q the number of latent variables in the model*, then:

- **β** (beta) is a QxQ matrix
  - Regression coefficients between latent vars
- **Ψ** (psi) is a QxQ matrix
  - (Co)variances of latent vars
- **Λ** (lambda) is a *P*xQ matrix
  - Factor loadings, relating observed to latent vars
- **Θ** (theta) is a *PxP* matrix
  - Measurement error (co)variances of observed vars

* and there are no formative latent variables and all regression relationships specified are between latent variables only

## Examples and exercises

- Example 3.3 and 3.4 - part II
- Additional Exercise 1a

## Parameter estimation

- Most often, parameter estimation in a SEM is performed by maximum likelihood (ML)
- Sometimes, ML estimates have closed form solutions, and can be calculated directly using a fomula
  - e.g., ML estimates for the population mean and variance:

$$\hat{\mu}_X = \bar{X} = \frac{1}{N}\sum_N X \qquad \hat{\sigma}_X^2 = \frac{1}{N}\sum_N (X-\bar{X})^2$$

- SEMs generally have a large number of parameters to be estimated, and an iterative procedure is more efficient to estimate the parameters
  - Therefore, output reports 'lavaan converged normally after … iterations'

## Parameter estimation and model fit

The outcome of the optimization process provides:

1. The ML estimates of the parameter values
2. The standard errors of the ML parameter estimates
   - Based on the 2nd order derivative of the likelihood function
   - With large sample sizes, the ratio of each estimated parameter to its standard error is approximately *z*-distributed
     - Gives a z- and p-value for each parameter in the output
3. The value of the likelihood function $F_{ML}$
   - Under the null hypothesis (i.e., the model-implied cov matrix is the true cov matrix in the population), -2 times the log-likelihood value at the final parameter estimates follows a chi-square distribution with *df* degrees of freedom
     - Allows for a statistical test of overall model fit when $df > 0$
     - When $df = 0$, the model always fits perfectly: likelihood = 1 and log(likelihood) = 0

## Assessing model fit

- Model fit should be evaluated in several ways:
  1. Overal model fit: assessed with model fit indices
  2. Individual parameter estimates
     - Parameter estimates substantial and statistically (in)significant where expected?
     - Are estimated parameter values plausible? E.g., expected sign of regression coefficients? Values as large or small as expected? E.g., |standardized factor loadings| > .30 ?
  3. Possible sources of misfit
     - Strikingly large residuals (co)variances or means?
     - Strikingly large modification index values?

## Assessing overal model fit

- Statistical test of model fit: $\chi^2$ (df)
  - Tests whether difference between the population and model-implied covariance matrix is zero
- In a SEM model, $\chi^2$ value quantifies difference between:
  - observed (sample) covariance matrix $\boldsymbol{S}$ and
  - model-implied (population) covariance matrix $\widehat{\boldsymbol{\Sigma}}$
    - $\chi^2$=0 if model fits perfectly, when $\widehat{\boldsymbol{\Sigma}} - \boldsymbol{S} = \boldsymbol{0}$
    - In all other cases, $\chi^2 > 0$
      - The larger the difference between $\widehat{\boldsymbol{\Sigma}}$ and $\boldsymbol{S}$, the larger the $\chi^2$ value

## Assessing overal model fit

- The larger the difference between $\widehat{\boldsymbol{\Sigma}}$ and $\boldsymbol{S}$, the larger the $\chi^2$ value, but:
- $\chi^2$ value is also affected by other factors, affecting type I and II error rates of the $\chi^2$ test:
  - Sample size
    - $\chi^2$ value almost always significant with sample sizes > 75
    - $\chi^2$ assesses statistical significance, but what about substantial significance?
    - One remedy: fit indices, are less dependent on sample size
  - Model complexity
    - More observed variables in model -> larger $\chi^2$ value
    - Remedy: Evaluate individual parameter estimates and residual (co)variances to assess model fit
  - Departures from multivariate normality
    - Increasing non-normality -> in- or deflated $\chi^2$ value
    - Remedy: use robust ML estimation

## Assessing overal model fit

- In addition to $\chi^2$(df), many other model fit indices
  - Lavaan provides > 40 of them for a single model
  - Have to make a selection:
    - Incremental fit indices (e.g., CFI)
    - Parsimony-based indices (e.g., RMSEA, AIC, BIC)
    - Absolute fit indices (e.g., SRMR)

## Incremental fit indices

- Higher values indicate better fitting model (range: 0-1; rarely, values > 1 occur)
- Compare the fit of the proposed model with that of a null model
  - The null model has:
    - Zero correlation between variables in the model (so no latent variables)
    - Variances of observed variables equal to sample variances
- Value depends on the average size of the correlations in the data
  - If average correlation between variables is not very high, then incremental fit indices not very high.

## Incremental fit indices

- Comparative fit index
  - Let d = $\chi^2 - df$
  - CFI = $\dfrac{\text{d(Null Model) - d(Proposed Model)}}{\text{d(Null Model)}}$
- Bentler-Bonett Index or Normed Fit Index (NFI)
  - $\dfrac{\chi^2\text{(Null Model) - }\chi^2\text{(Proposed Model)}}{\chi^2\text{(Null Model)}}$
  - Not so often used, due to no penalty for model complexity
- Tucker Lewis Index or Non-normed Fit Index (NNFI):
  - $\dfrac{\chi^2\text{/df(Null Model) - }\chi^2\text{/df(Proposed Model)}}{\chi^2\text{/df(Null Model) - 1}}$

## Parsimony-based indices

- Information-theoretic criteria:
  - Model with lowest value has best fit
  - Note that there are various ways to calculate AIC, so never compare between software packages!
- AIC: Akaike's Information Criterion
  - Penalty for every additional, freely estimated parameter is 2
- BIC: Bayesian Information Criterion
  - Penalty for every additional, freely estimated parameter is nat.log(N), where N is the total sample size
- SSABIC: Sample-Size Adjusted BIC
  - Penalty for every additional, freely estimated parameter is ln([N+2]/24)

## Parsimony-based indices

- RMSEA: Root Mean Square Error of Approximation

$$RMSEA = \sqrt{\frac{\chi^2 - df}{df \cdot (N-1)}}$$

  - Lower values indicate better fitting model
    - Also, confidence interval can be calculated
    - And the p-value for RMSEA <= 0.05 (if p-value > .05, hypothesis of close fit is retained)
- $\chi^2$/df ratio
  - Smaller values indicate better fit
  - Various rules of thumb have been proposed, ranging from 2 to 6 (what is good depends also on sample size)

## Absolute fit indices

- SRMR: Standardized Root Mean Squared Residual

$$RMR = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}(s_{ij} - \hat{\sigma}_{ij})^2}{p(p+1)/2}}$$

$s_{ij}$ is an element of the empirical covariance matrix S,

$\hat{\sigma}_{ij}$ is an element of the model-implied matrix covariance $\Sigma(\hat{\theta})$, and

$p$ is the number of observed variables.

  - Average difference between the observed and model-implied correlations
  - Has no penalty for model complexity
  - SRMR = 0 indicates perfect fit

## Overall model fit – cut-off values

- Based on simulations, Hu & Bentler (1999) derived the following cut-off values for <u>good</u> model fit:
  - CFI/TLI ≥ .95
  - SRMR ≤ .08
  - RMSEA ≤ .06
- Other authors suggest more lenient criteria
  - Sometimes, CFI ≥ .90 and/or RMSEA ≤ .08 called 'adequate' or 'acceptable'
- Model fit is not an all-or-nothing question, rules-of-thumb above offer a good starting point

## Examples and exercises

- Example 3.3 and 3.4 – part III
- Exercise 3.1
- Exercise 3.2

## Improving model fit

- **Residual (co)variances**
  - Observed sample (co)variances minus model-implied covariances
  - Can be obtained in lavaan with the residuals() function
  - Using this information, the model may be improved

## Improving model fit

- **Modification indices**
- Give an estimate of how much the $\chi^2$-value of model fit will decrease when a parameter is freely estimated
- It can be interpreted as a $\chi^2$-value with 1 df
  - Rule of thumb: if MI > 5, consider estimating parameter freely

**Percentage Points of the Chi-Square Distribution**

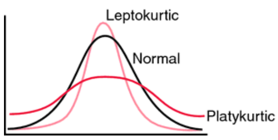
| Degrees of Freedom | Probability of a larger value of $x^2$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0.99 | 0.95 | 0.90 | 0.75 | 0.50 | 0.25 | 0.10 | 0.05 | 0.01 |
| 1 | 0.000 | 0.004 | 0.016 | 0.102 | 0.455 | 1.32 | 2.71 | 3.84 | 6.63 |
| 2 | 0.020 | 0.103 | 0.211 | 0.575 | 1.386 | 2.77 | 4.61 | 5.99 | 9.21 |
| 3 | 0.115 | 0.352 | 0.584 | 1.212 | 2.366 | 4.11 | 6.25 | 7.81 | 11.34 |
| 4 | 0.297 | 0.711 | 1.064 | 1.923 | 3.357 | 5.39 | 7.78 | 9.49 | 13.28 |
| 5 | 0.554 | 1.145 | 1.610 | 2.675 | 4.351 | 6.63 | 9.24 | 11.07 | 15.09 |

## Examples and exercises

- Example 3.3 and 3.4 - parts IV and V
- Additional exercise 1c
- Additional exercise 2

## Robust ML estimation

- Robust ML estimation, like ML, assumes the data follow a multivariate normal distribution, but that the data have more or less kurtosis than a normal distribution
  - Thus does not correct for skewness!
- Kurtosis: measure of the shape of the distribution
  - From Greek word for bulging
- The degree of kurtosis in a data set is related to how incorrect the log-likelihood value will be
  - Leptokurtic data: $\chi^2$ too large, SEs too small
  - Platykurtic data: $\chi^2$ too small, SEs too large




## Robust ML estimation

- **Parameter estimates** under MLR are just ML estimates
- SEs and model $\chi^2$ value are adjusted under MLR, depending on kurtosis of data:
  - Model $\chi^2$ value and associated fit statistics are adjusted
    - smaller $\chi^2$ when data are leptokurtic
    - larger $\chi^2$ when data are platykurtic
  - Model SEs are adjusted
    - smaller SEs when data are leptokurtic
    - larger SEs when data show platykurtosis
- If data have normal kurtosis, no adjustment is made (so safe to always use MLR)

## Robust ML estimation

- Invoked by adding argument 'estimator = "MLR" ' in model-fitting function (e.g., lavaan(), sem(), cfa(), growth functions)
- Works only when raw data is supplied
  - When only covariance matrix (and/or means) are supplied, there is no info about the kurtosis of the data, so adjusted the standard errors and test statistic is not possible

## Examples and exercises

- Additional exercise 3