

Confirmatory Factor Analysis

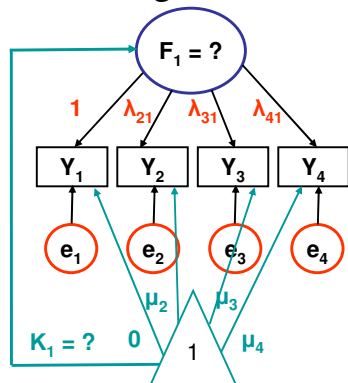
Part 2: Model Evaluation

- Today's topics:
 - Review of Factor Model Identification
 - 4 Steps in Model Evaluation
 - Model Fit and Fit Indices
 - Making Sense of the Model
 - Troubleshooting
 - Wrapping Up...

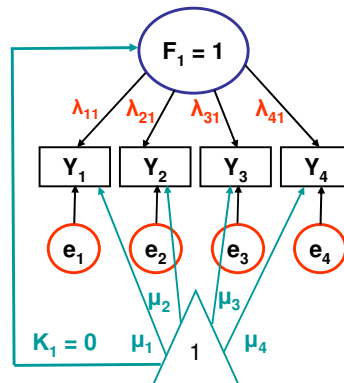
Factor Model Identification

- Goal: *Name that Tune* → Reproduce observed covariance matrix among items with as few estimated parameters as possible
 - Maximum likelihood usually used to estimate model parameters
 - **Measurement Model:** Factor loadings, item intercepts, error variances
 - **Structural Model:** Factor variances and covariances, factor means
 - Global model fit is evaluated as difference between model-predicted matrix and observed matrix (but only the covariances really contribute)
- How many possible parameters can you estimate (total DF)?
 - **Total DF depends on # ITEMS** → p (NOT on # people)
 - Total number of 'unique elements' in covariance matrix
 - Unique elements = each **variance**, each **covariance**, each **mean**
 - Total unique elements = $(p(p+1) / 2) + p$ → if 4 items, then $((4*5)/2) + 4 = 14$
- Model degrees of freedom (df)
 - Model df = # possible parameters – # estimated parameters

CFA Model Identification: *Scaling the Factor Mean and Variance*



“**Marker Item**” → Fix 1 item intercept to 0, loading to 1
Estimate factor mean and variance



“**Z-Score**” → Fix factor mean to 0 and variance to 1
Estimate all intercepts and loadings

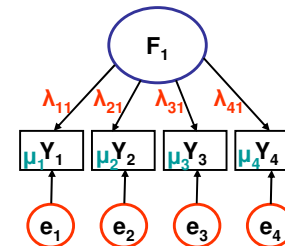
Over-Identified Factor: 4+ Items

- Model is over-identified when there are fewer unknowns than pieces of information with which to estimate them
 - Parameter estimates have a unique solution that will NOT perfectly reproduce the observed matrix
 - **NOW we can test model fit**

Total possible df = unique elements = 14

0 factor variances	1 factor variance
0 factor means	1 factor mean
4 loadings	OR 3 item loadings
4 item intercepts	3 item intercepts
4 error variances	4 error variances

$$df = 14 - 12 = 2$$



Did we do a 'good enough' job reproducing the matrix with 2 fewer parameters than was possible to use?

CFA Model Equations with Item Intercepts

- Measurement model per item (numbered) for subject s :
 - $Y_{1s} = \mu_1 + \lambda_{11}F_{1s} + 0F_{2s} + e_{1s}$
 - $Y_{2s} = \mu_2 + \lambda_{21}F_{1s} + 0F_{2s} + e_{2s}$
 - $Y_{3s} = \mu_3 + \lambda_{31}F_{1s} + 0F_{2s} + e_{3s}$
 - $Y_{4s} = \mu_4 + \lambda_{41}F_{1s} + 0F_{2s} + e_{4s}$
 - $Y_{5s} = \mu_5 + 0F_{1s} + \lambda_{52}F_{2s} + e_{5s}$
 - $Y_{6s} = \mu_6 + 0F_{1s} + \lambda_{62}F_{2s} + e_{6s}$
 - $Y_{7s} = \mu_7 + 0F_{1s} + \lambda_{72}F_{2s} + e_{7s}$
 - $Y_{8s} = \mu_8 + 0F_{1s} + \lambda_{82}F_{2s} + e_{8s}$
- You decide how many factors and whether each item loads (loading then estimated) or not.
- Unstandardized loadings (λ) are the slopes of regressing the response (Y) on the factor (X).
- Standardized loadings are the slopes in a correlation metric (and Std Loading² = reliability).
- Intercepts (μ) are expected value of Y (item) when all factors (X 's) are 0 (no misfit).
- The equation predicting each item resembles a linear regression model:
- $$Y_{is} = B_{0i} + B_{1i}X_{1s} + B_{2i}X_{2s} + e_{is}$$

Estimation via Maximum Likelihood

- Goal is to find the parameter estimates (loadings, error variances, factor variances, and factor covariances) that make the observed data matrix most likely
 - Program iterates until answers (estimates) don't change much anymore – this point is called 'model convergence'
 - Start values help achieve convergence in complicated models
- Assumptions of ML:
 - Is asymptotic → correct estimates given large sample size
 - Item responses can be treated as continuous, interval variables
 - Items have multivariate normal distribution
 - Non-normality → bias in SE and obtained model fit index of χ^2
 - Other possible estimators, too (more later on this)

The Big Picture of Model Fit

- Aspects of observed data to be modeled (*assuming a z-score metric for the factor for simplicity*):
- Model equation: $Y_{is} = \mu_i + \lambda_i F_s + e_{is}$
 - **Mean** per item: Represented via intercept μ per item
 - Not a source of misfit (constraints rarely applied on intercepts)
 - **Variance** per item: Represented as weighted (F) + (e)
 - $\text{Var}(Y_i) = (\lambda_i^2) \cdot \text{Var}(F) + \text{Var}(e_i) \rightarrow$ note imbalance of λ and e^2
→ output is given as λ and $\text{Var}(e)$
 - Factor and error variances are additive → not a source of misfit (whatever F doesn't get, e picks up to get back to total Y variance)
 - **Covariance** among items: Represented via factor loading λ_i
 - Loadings multiplied predict what observed covariance should be... but they may not be right → **THE SOURCE OF MISFIT**

4 Steps in Assessing Model Fit

1. Global model fit
 - *Does the model 'work' as a whole?*
2. Local model fit
 - *Are there any more specific problems?*
3. Inspection of model parameters
 - *Do the numbers make sense? Are they useful?*
4. Reliability and information per item
 - *How 'good' is my test? How 'good' is each item?*

The Basis of Global Model Fit

- Assess global model fit via maximum likelihood
 - How well did the model reproduce the observed matrix of variances and covariances among the p items?
 - Σ = model-predicted matrix (not a sum this time, I know, sorry)
 - S = sample observed matrix
- Fitting function minimizes difference between Σ and S
 - $F_{ML} = \ln|S| - \ln|\Sigma| + \text{trace} [(S)(\Sigma^{-1})] - p$

Determinant example:	
a b	= ad-bc
c d	

 - $\ln|S|$ = natural log of determinant of S
 - $\ln|\Sigma|$ = natural log of determinant of Σ
 - $\text{trace} [(S)(\Sigma^{-1})]$ = sum of diagonals of result of S divided by Σ
 - If S and Σ match perfectly, their difference will be 0
 - If so, $[(S)(\Sigma^{-1})]$ will be an identity matrix, with sum equal to p

Indices of Global Model Fit

- Primary: obtained model $\chi^2 = F_{ML}(N-1)$
 - χ^2 is evaluated based on model df (# parameters left over)
 - Tests null hypothesis that $\Sigma = S$ (that model is perfect), so significance is undesirable (smaller χ^2 , bigger p-value is better)
 - Just using χ^2 is insufficient, however:
 - Distribution doesn't behave like a true χ^2 if sample sizes are small or if items are non-normally distributed
 - Obtained χ^2 depends largely on sample size
 - Is unreasonable null hypothesis (perfect fit??)
- Because of these issues, alternative measures of fit are usually used in conjunction with the χ^2 test of model fit
 - Absolute Fit Indices (besides χ^2)
 - Parsimony-Corrected; Comparative (Incremental) Fit Indices

Indices of Global Model Fit

- Absolute Fit: χ^2
 - Don't use 'ratio rules' like $\chi^2/df > 2$ or $\chi^2/df > 3$
- Absolute Fit: **SRMR**
 - **Standardized Root Mean Square Residual**
 - Get difference of Σ and S \rightarrow residual matrix
 - Sum the squared residuals in matrix, divide by number of residuals summed
 - Ranges from 0 to 1: smaller is better
 - “.08 or less” \rightarrow good fit
- See also: **RMR (Root Mean Square Residual)**

Indices of Global Model Fit

- Parsimony-Corrected: **RMSEA**
 - **Root Mean Square Error of Approximation**
 - Relies on a non-centrality parameter (NCP)
 - Indexes how far off your model is \rightarrow χ^2 distribution shoved over
 - $NCP \rightarrow d = (\chi^2 - df) / (N-1)$ Then, $RMSEA = \text{SQRT}(d/df)$
 - RMSEA ranges from 0 to 1; smaller is better
 - $< .05$ or $.06$ = “good”, $.05$ to $.08$ = “acceptable”, $.08$ to $.10$ = “mediocre”, and $> .10$ = “unacceptable”
 - In addition to point estimate, get 90% confidence interval
 - RMSEA penalizes for model complexity – it's discrepancy in fit per df left in model (but not sensitive to N, although CI can be)
 - Test of “close fit”: null hypothesis that $RMSEA \leq .05$

Indices of Global Model Fit

Comparative (Incremental) Fit Indices

- Fit evaluated relative to a 'null' model (of 0 covariances)
- Relative to that, your model should be great!
- **CFI: Comparative Fit Index**
 - Also based on idea of NCP ($\chi^2 - df$)
 - $CFI = 1 - \frac{\max[(\chi^2_T - df_T), 0]}{\max[(\chi^2_T - df_T), (\chi^2_N - df_N), 0]}$ T = target model
N = null model
 - From 0 to 1: bigger is better, > .90 = "acceptable", > .95 = "good"
- **TLI: Tucker-Lewis Index (= Non-Normed Fit Index)**
 - $TLI = \frac{(\chi^2_N/df_N) - (\chi^2_T/df_T)}{(\chi^2_N/df_N) - 1}$
 - From <0 to >1, bigger is better, >.95 = "good"

4 Steps in Model Evaluation

1. Assess global model fit

- Recall that item intercepts, factor means, and variances are just-identified → *misfit comes from messed-up covariances*
- χ^2 is sensitive to large sample size
- Pick at least one global fit index from each class; hope they agree (e.g., CFI, RMSEA)
- If model fit is not good, you should NOT be interpreting the model estimates
 - They will change as the model changes
- If model fit is not good, it's your job to find out WHY
- If model fit is good, it does not mean you are done, however... proceed to step 2

4 Steps in Model Evaluation

2. Identify localized model strain

- Global model fit means that the observed and predicted matrices aren't too far off on the whole... says nothing about the specific matrix elements
- Should inspect **normalized model residuals** for that
 - Available via RESIDUAL option in Mplus
 - Normalized as residual/SE → like a z-score
 - Anything bigger than 2-3ish indicates "localized strain"
 - Positive residual → More related than you predicted
 - More than just the factor creating a covariance
 - Negative residual → Less related than you predicted
 - Not as related as you said they should be
- **Evidence of localized strain tells you where the problems are, but not what to do about them...**

4 Steps in Model Evaluation

2. Identify localized model strain, continued...

- Another approach: **Modification Indices (aka, voo-doo)**
- Two flavors: What to Add, and What to Drop
 - LaGrange Multiplier: How much the χ^2 would decrease by adding a particular model parameter (e.g., cross-loading, error correlation)
 - Usually only pay attention if > 4 for 1 df
 - Get expected parameter estimate for what's to be added – only pay attention if its effect size is meaningful
 - Only pay attention if you can INTERPRET AND DEFEND IT
 - Wald Test: How much χ^2 would increase by dropping a particular model parameter (e.g., factor loading, error covariance)
 - Make sure you aren't dropping anything "important"
- Implement these ONE AT A TIME, because one change can alter the rest of the model substantially

'Fixing' the Model

- A common source of misfit is due to items that remain too correlated after accounting for their common factor
- Solutions for this:
 - Add **error covariance** (i.e., as suggested by voo-doo indices)
 - Is additive: $Cov(y_1, y_2) = cov \text{ due to } F + cov \text{ due to error covariance}$
 - **Error covariances are unaccounted for multi-dimensionality**
 - this means you have measured your factor + something else that those items have in common (e.g., stem, valence, specific content)
 - Just one or two problematic pairings (too correlated)?
 - Do you need both items? Try a reduced model without one of them. However – models with differing # items are NOT COMPARABLE. You'd need to keep the item but drop the loading for a nested model.
 - Lots of problematic pairings? **Re-consider dimensionality.**
 - I'd recommend against invoking cross-loadings.

Psyc 948 Class 5a
17 of 28

Messing with the Model

- Can assess whether changing the model (adding or subtracting parameters) impacts the model fit:
 - **Nested models can be compared with χ^2 difference tests**
 - Step 1: Calculate difference of χ^2_{old} and χ^2_{new}
 - Step 2: Calculate difference in df_{old} and df_{new}
 - Compare χ^2_{diff} on $df = df_{diff}$ to critical values table (or excel CHIDIST)
 - Add 1 parameter? $\chi^2_{diff} > 3.84$, add 2: $\chi^2_{diff} > 5.99...$
 - If **adding** a parameter, the model can either stay the same OR get **better**, as indexed by the χ^2
 - If **removing** a parameter, the model can either stay the same OR get **worse**, as indexed by the χ^2
 - If testing parameters that can't be negative (like variances = 0?), then should use $p < .10$ instead of $p < .05$ (or mixture χ^2 tables)

Psyc 948 Class 5a
18 of 28

Messing with the Model

- If the to-be-compared models are non-nested, the χ^2 difference test is not applicable
 - e.g., should Y_1 load on F_2 instead of F_1 ?
- Use AIC (Akaike Information Criteria) or BIC instead
 - Is NOT a significance test – is "evidence"
 - Smaller is better, but there are no critical values
- For both nested or non-nested model comparisons, differences in other fit indices should be examined, too
 - χ^2_{diff} still sensitive to sample size in comparing models
 - No real critical values for changes in other fit indices, however
 - They may disagree (especially RMSEA, which likes parsimony)

Psyc 948 Class 5a
19 of 28

4 Steps in Model Evaluation

3. Inspect **parameter effect sizes** and significance levels
 - Both of these correlation matrices will have perfect fit for a one-factor model:

	Y1	Y2	Y3	Y4
Y1	1			
Y2	.1	1		
Y3	.1	.1	1	
Y4	.1	.1	.1	1

Loading = $\text{SQRT}(.1) \approx .32$

	Y1	Y2	Y3	Y4
Y1	1			
Y2	.8	1		
Y3	.8	.8	1	
Y4	.8	.8	.8	1

Loading = $\text{SQRT}(.8) \approx .89$

- **Model fit does not guarantee meaningful factor loadings**

Psyc 948 Class 5a
20 of 28

4 Steps in Model Evaluation

3. Inspect **parameter effect sizes** and significance levels
 - Model fit does not guarantee meaningful factor loadings
 - Can reproduce lack of covariance quite well and still not have anything useful – e.g., factor loading of .2 → 4% shared variance?
 - Get SEs and p-values for unstandardized (and standardized) estimates (usually report SE from unstandardized)
 - Marker indicators (set to 1 for identification) won't have significance tests for their loadings because they are fixed at 1, but you'll still get a standardized factor loading test for them
 - Standardized loadings help to judge relative importance
 - Make sure all estimates are within bounds
 - No standardized factor loadings > 1 (unless the indicator has cross-loadings, in which case this is actually possible)
 - No negative factor variances or negative error variances

Psyc 948 Class 5a
21 of 28

4 Steps in Model Evaluation

4. Calculate item information and model-based reliability
 - **Item Information** = $(\text{unstandardized } \lambda)^2 / (e^2)$
 - What proportion of item variance is “true” relative to error?
 - Examining the unstandardized loadings by themselves is not enough, as their relative contribution depends on how much error variance the item has.
 - The standardized loadings will give you the same rank order in terms of item information, which is why information is not often used within CFA (but stay tuned for information in IRT).
 - **“Omega” Test Reliability** = $(\Sigma\lambda)^2 / [(\Sigma\lambda)^2 + \Sigma(e^2)]$
 - Squared sum of ***standardized*** factor loadings, over that + summed error variances
 - Although our CFA book recommends unstandardized, I've found a peculiarity such that omega can differ across methods of model identification. Because the standardized solution is always the same, I'd recommend using that version of the loadings instead.

Psyc 948 Class 5a
22 of 28

Testing CTT Assumptions in CFA

- **Alpha** is reliability assuming two things:
 - All factor loadings (discriminations) are equal, or that the items are ‘true-score (tau) equivalent’
 - That **unidimensionality** holds (which we now test within factor models)
- We can test the assumption of **tau-equivalence** too via nested model comparisons in which the loadings are constrained to be equal – does model fit decrease?
 - If so, don't use alpha – use model-based reliability (omega) instead. Omega assumes unidimensionality, but not tau-equivalence.
 - Research has shown alpha can be an over-estimate or an under-estimate depending on particular data characteristics.
- The assumption of ‘**Parallel items**’ is then testable by constraining error variances to be equal, too – does model fit decrease?
 - ‘Parallel items’ will hardly ever hold in real data.
 - Note that if tau-equivalence doesn't hold, then neither does ‘parallel’.

Psyc 948 Class 5a
23 of 28

What can go wrong?

- Considering the number of factors...
 - Factors correlated > .85 may suggest a simpler structure
 - Nested model comparison: test correlation fixed at 1 → is fit hurt?
 - Do you need additional method factors or error covariances?
- Considering factor loadings...
 - No-loadings: If the item isn't related, it isn't measuring the construct, and you probably don't need it
 - Wrong-loadings: Could happen, but not too likely
 - Negative loadings: Reverse-code first, so these shouldn't happen
 - Cross-loadings: If the item measures more than one thing, it probably isn't going to be useful (and will complicate interpretation)
- Considering error correlations...
 - Can you defend them? Did you expect them? Will they replicate?

Psyc 948 Class 5a
24 of 28

What can go wrong?

- Error message: “non-positive definite”
 - Both observed and predicted matrices must be positive definite
 - “Non-Positive Definite” means that the determinant is approaching 0, or that the matrix is singular
 - Some variables are (nearly) perfectly correlated
 - Double-check that data are being read in correctly; otherwise you may need to drop indicators that are too highly correlated
- Structural under-identification
 - Does every factor have a metric and at least 3 items?
 - Does the marker item actually load on the factor???
- Empirical under-identification
 - More likely with smaller sample sizes, fewer indicators per factor, and indicators with low communalities (R^2 values)

Psyc 948 Class 5a
25 of 28

Open in case of emergency...

- If good model fit seems hopeless, you may need to go back to the drawing board... almost
- Brown suggests an “E/CFA” approach of estimating an exploratory-like model within a CFA framework:
 - Fix each factor variance to 1
 - Each factor gets one indicator that only loads on it (fixed to 1)
 - Rest of indicators load on all factors
 - Why bother? To get significance tests of factor loadings
 - May suggest a useful alternative structure, which should then ideally be replicated in an independent sample using CFA

Psyc 948 Class 5a
26 of 28

Model Fit: Summary

- The primary advantage of working in a CFA framework is obtaining indices of global and local model fit
 - X^2 and assorted model fit indices indicate how well the model-predicted covariance matrix matches the observed data matrix...
 - .. But normalized residuals should still be examined for evidence of ‘local misfit’ (e.g., between certain items)
 - Nested model comparisons can be conducted in order to improve the fit of the model or to simplify the model...
 - ... But careful relying too heavily on modification indices to do so
 - Size and significance of model parameters matters, too
 - ... How well are your factors really defined anyway?
 - Watch out for out-of-bound estimates – means something is wrong

Psyc 948 Class 5a
27 of 28

The Big Picture of CFA

- **CFA unit of analysis is the ITEM: $Y_{is} = \mu_i + \lambda_i F_s + e_{is}$**
 - **Linear** regression relating continuous Y to latent predictor F
 - Both items AND subjects matter in predicting responses
 - Factors are estimated as separate entities based on the observed covariances among items – factors represent testable assumptions
 - Items are unrelated after controlling for factors → local independence
- **Because item responses are included:**
 - Items are allowed to vary in discrimination (factor loadings)
 - thus, exchangeability (tau-equivalence) is a testable hypothesis
 - Because difficulty (item intercepts) do not contribute to the covariance, they don't really matter in CFA (unless testing factor mean differences)
 - To make a test better, you need more items
 - **What kind of items? Ones with greater information (λ^2/e^2)**
 - Measurement error is still assumed constant across the latent trait
 - **People low-medium-high in Factor Score are measured equally well**

Psyc 948 Class 5a
28 of 28