Confirmatory Factor Analysis Part 1: Model Identification

- Today's topics:
  - Comparison of EFA and CFA
  - CFA Model
    - Parameters and Model Predictions
    - Model Identification
    - Components vs. Factors
    - Factor Scores
  - Next time: Model Fit Evaluation

EFA vs. CFA

- What gets analyzed…
  - **EFA: Correlation matrix (of items/indicators)**
    - Correlations among observed indicators are relevant
    - Only a standardized solution, so the means and variances of the observed indicators are irrelevant
  - **CFA: Covariance matrix (of items/indicators)**
    - Variances and covariances of observed items are relevant
    - Means are often ignored
      - Without, assume each item is mean-centered
      - We will include means (actually, *intercepts*) per item
    - Get unstandardized AND standardized solutions
      - Each is important for different reasons

EFA vs. CFA, continued

- How we get an interpretable solution…
  - **EFA: Rotation**
    - All items load on all factors
    - Goal is to pick a rotation that gives closest approximation to simple structure (clear factors, fewest cross-loadings)
    - No way of separating 'content' from 'method' factors
  - **CFA: Your job in the first place!**
    - CFA must be theory-driven
    - You specify number of factors and their inter-correlations
    - You specify which items load on which factors (yes/no)
    - You specify any unique (error) relations for method variance

EFA vs. CFA, continued

- How we judge model fit…
  - **EFA: Eye-balls and Opinion**
    - #Factors? Scree-ish plots, interpretability…
    - Which rotation? Whichever makes most sense…
    - Which indicators load? Cut-off of .3-.4ish
  - **CFA: Inferential tests via Maximum Likelihood (ML)**
    - Global model fit test
    - Significance of item loadings
    - Significance of error variances (and covariances)
    - Ability to test model constraints via change in model fit

EFA vs. CFA, continued

- What we do with the latent factors…
  - **EFA: Compute factor scores if you must…**
    - Factor indeterminacy issues
    - Inconsistency in how factor models are applied to data
      - Factor model based on common variance only
      - Summing items? That's using total variance (component)
  - **CFA: Let them be part of the model**
    - Don't need individual factor scores
    - Test relations with latent factors directly through SEM
      - Factors can either be predictors (exogenous) or outcomes (endogenous) as needed
      - Relationships will be disattenuated for measurement error

Confirmatory Factor Analysis (CFA)

- **In CFA the ITEM (indicator) is the unit of analysis:**
  $\mathbf{Y_{is} = \mu_i + \lambda_i F_s + e_{is}}$ → both items AND subjects matter
  - Observed response for item *i* and subject *s*
    = intercept of item *i* (μ)
    + subject *s*'s latent trait/factor (F), item-weighted by λ
    + error (e) of item *i* and subject *s*

- **What does this look like? Linear regression (without a real X)!**
  - $Y_s = B_0 + B_1F_s + e_s$ → written for each item → $\mathbf{Y_{is} = B_{0i} + B_{1i}F_s + e_{is}}$
  - Intercept $B_{0i} = \mu_i =$
  - Slope of Factor $B_{1i} = \lambda_i =$
  - Residual $e_{is} = e_{is} =$

2 Types of CFA Solutions

- Just like linear regression comes in unstandardized and standardized flavors, so does CFA output:
- <u>Unstandardized</u> → predicts scale-sensitive original item response:
  - $\mathbf{Y_{is} = \mu_i + \lambda_i F_s + e_{is}}$
  - *Useful when comparing solutions across groups or time*
  - Note the solution asymmetry: item parameters μ_i and λ_i will be given in the item metric, but $\mathbf{e_{is}}$ will be given as the error variance across persons for that item
  - $Var(Y_i) = [\mathbf{\lambda_i^2} * Var(F)] + Var(e_i)$
- <u>Standardized</u> → STDYX solution transformed to $Var(Y_i)$=1, Var(F)=1:
  - *Useful when comparing items within a solution*
  - Standardized intercept = $\mathbf{\mu_i}$ / **SD(Y)** → not typically reported
  - Standardized factor loading = **[λ_i * SD(F)] / SD(Y) = item correlation with factor**
  - Standardized error variance = 1 – standardized λ_i^2 = "variance due to *not* factor"
  - R^2 for item = standardized λ_i^2 = "variance due to the factor"

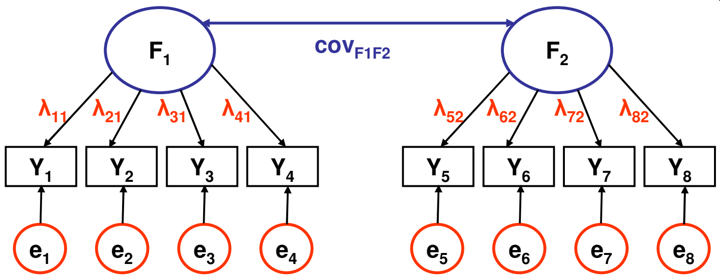
Assumptions of CFA Models

- **Dimensionality** is assumed known (often 1 latent trait)
  - Local Independence → e's are independent after controlling for factor(s)
- **Linear model** → a one-unit change in latent trait/factor F has same increase in expected item response (Y) at all points of factor (X)
  - Won't work well for binary data… thus, we need IRT
  - Sometimes of questionable utility for Likert scale data
- Goal is to **predict covariance** between items → basis of model fit
  - Variance will always be perfectly reproduced; covariance will not be
- CFA models are **usually presented without μ_i** (the item intercept)
  - μ_i doesn't really matter in CFA because it doesn't contribute to covariance, but we will keep it for continuity with IRT
  - Item intercepts are also important when dealing with factor mean diffs

CFA Model Parameters

(no Factor Means or Item Intercepts)

(But some of these values will have to be restricted for the model to be identified.)



Measurement Model:

- λ's = factor loadings
  - subscripts = item, factor
- e's = error variances
  - errors are all uncorrelated

Structural Model:

- F's = factor variances
- Cov = factor covariances
  - Will often be larger than in EFA (no cross-loadings)

CFA Model Predictions

F_1 BY Y_1-Y_4, F_2 BY Y_5-Y_8

Two items from same factor (room for misfit):

- Unstandardized solution: $Covariance_{y1,y4}$ = λ_{11}*$Var(F_1)$*λ_{41}
- Standardized solution: $Correlation_{y1,y4}$ = λ_{11}* (1)*λ_{41} → **std loadings**
- ONLY reason for $cor_{y1,y4}$ is common factor (local independence, LI)

Two items from different factors (room for misfit):

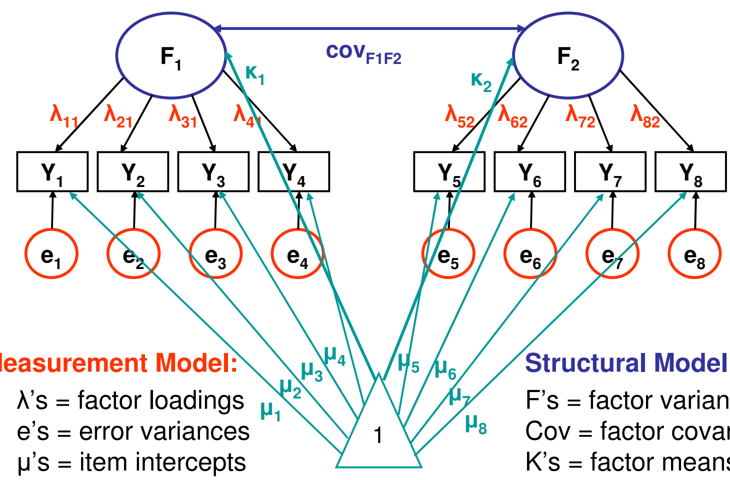
- Unstandardized solution: $Covariance_{y1,y8}$ = λ_{11}*$cov_{F1,F2}$*λ_{82}
- Standardized solution: $Correlation_{y1,y8}$ = λ_{11}*$cor_{F1,F2}$*λ_{82} → **std loadings**
- ONLY reason for $cor_{y1,y8}$ is correlation between factors (again, LI)

Variances are additive (and will be reproduced correctly):

- **$Var(Y_1) = (\lambda_{11}^2)*Var(F_1) + Var(e_1)$ → note imbalance of λ^2 and e**

CFA Model WITH Factor Means and Item Intercepts

(But some of these values will have to be restricted for the model to be identified.)



Measurement Model:

λ's = factor loadings
e's = error variances
μ's = item intercepts

Structural Model:

F's = factor variances
Cov = factor covariances
K's = factor means

CFA Model Equations with Item Intercepts

- Measurement model per item (numbered) for subject *s*:
  - $Y_{1s} = \mu_1 + \lambda_{11}F_{1s} + 0F_{2s} + e_{1s}$
  - $Y_{2s} = \mu_2 + \lambda_{21}F_{1s} + 0F_{2s} + e_{2s}$
  - $Y_{3s} = \mu_3 + \lambda_{31}F_{1s} + 0F_{2s} + e_{3s}$
  - $Y_{4s} = \mu_4 + \lambda_{41}F_{1s} + 0F_{2s} + e_{4s}$
  - $Y_{5s} = \mu_5 + 0F_{1s} + \lambda_{52}F_{2s} + e_{5s}$
  - $Y_{6s} = \mu_6 + 0F_{1s} + \lambda_{62}F_{2s} + e_{6s}$
  - $Y_{7s} = \mu_7 + 0F_{1s} + \lambda_{72}F_{2s} + e_{7s}$
  - $Y_{8s} = \mu_8 + 0F_{1s} + \lambda_{82}F_{2s} + e_{8s}$

The equation predicting each item resembles a linear regression model:

$Y_{is} = B_{0i} + B_{1i}X_{1s} + B_{2i}X_{2s} + e_{is}$

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 Loading^2 = reliability).

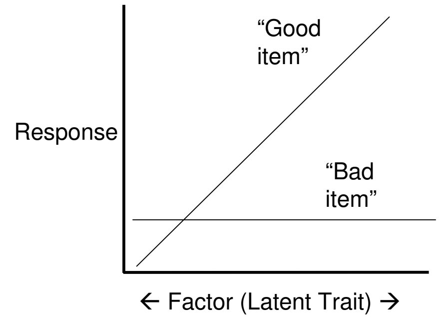
Intercepts (μ) are expected value of Y (item) when all factors (X's) are 0 (no misfit).

Revisiting Vocabulary: Item Psychometric Properties

- **Discrimination**: How related each item is to the latent trait
  - In CTT, discrimination is indexed by the item-total correlation
    - The total score is the best estimate of the latent trait in CTT
  - In **CFA**, discrimination is indexed by the **factor loading (λ)**
    - We now have a factor estimated by using covariance among items
  - In IRT, discrimination will be indexed by the slope of the item characteristic *curve* (no longer a linear model)

- **Difficulty**: Location of item on the latent trait metric
  - In CTT, difficulty is indexed by the item mean (for quantitative items) or proportion passing (for binary items)
  - In **CFA**, difficulty is indexed by the **item intercept (μ)**
  - In IRT, difficulty will indexed by the item's location (and will be much more important than it is in CFA models)

Why Item Intercepts Are Often Ignored…



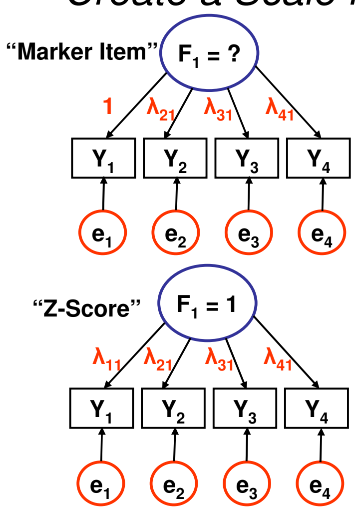
If an item is "good" (strong slope between factor and item response → strong loading, it is equally good across the latent trait.

If an item is "bad", it is equally bad across the latent trait.

Further, the model is based on accounting for covariance – the means fall out of that equation.

However – item intercepts will be important for testing hypotheses about factor mean differences.

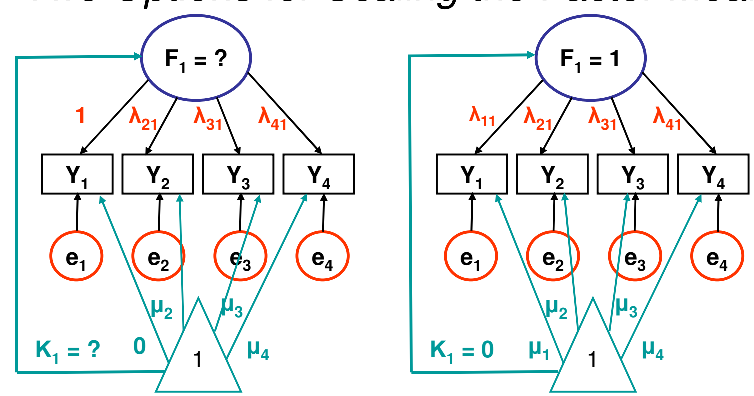
CFA Model Identification:
Create a Scale for the Pretend Variable



- The factor doesn't exist, so **it needs a scale** (a mean and variance):
- Two **equivalent** options to do so

- Create a scale for the VARIANCE:
  - **1) Scale with a marker item**
    - Fix one loading to 1; factor is scaled as reliable part of that marker item
    - Loading = .9, variance =16? $Var(F_1) = (.9^2)*16 = 12.96$
  - **2) Fix factor variance to 1**
    - Factor is interpreted as z-score
    - Can't be used in other models with higher-order factors (stay tuned for part 3)

CFA Model Identification:
Two Options for Scaling the Factor Mean



"Marker Item" → Fix 1 item intercept to 0; estimate factor mean

Item intercept is expected outcome when factor = 0 (when item = 0)

"Z-Score" → Fix factor mean to 0, estimate all item intercepts

Item intercept is expected outcome when factor = 0 (when item = mean)

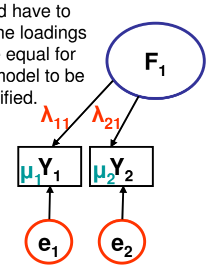
Factor Model Identification

- 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

Under-Identified Factor: 2 Items

- Model is <u>under-identified</u> when there are more unknowns then pieces of information with which to estimate them
  - Cannot be solved because there are an infinite number of different parameter estimates that would result in perfect fit
  - Example: Solve x + y = 7 ??

You'd have to set the loadings to be equal for the model to be identified.



Assumptions required to calculate reliability in CTT are due to under-identification.

Total possible df = unique elements = **5**

0 factor variances		**1 factor variance**
0 factor means		**1 factor mean**
2 loadings	**OR**	**1 item loading**
2 item intercepts		**1 item intercept**
2 error variances		**2 error variances**

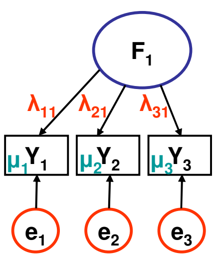
df = 5 – 6 = -1

If $r_{y1,y2} = .64$, then:

$\lambda_{11} = .800$, $\lambda_{21} = .800$??
$\lambda_{11} = .900$, $\lambda_{21} = .711$??
$\lambda_{11} = .750$, $\lambda_{21} = .853$??

Just-Identified Factor: 3 Items

- Model is <u>just-identified</u> when there are as many unknowns as pieces of information with which to estimate them
  - Parameter estimates have a unique solution that will perfectly reproduce the observed matrix
  - Example: Solve x + y = 7, 3x – y = 1



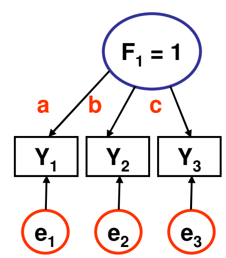
Total possible df = unique elements = **9**

0 factor variances		**1 factor variance**
0 factor means		**1 factor mean**
3 loadings	**OR**	**2 item loadings**
3 item intercepts		**2 item intercepts**
3 error variances		**3 error variances**

df = 9 – 9 = 0

Not really a model – more like a description

Solving a Just-Identified Model



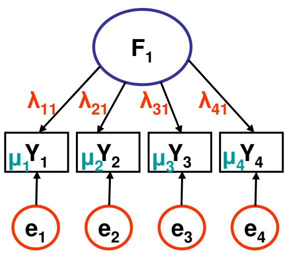
	Y_1	Y_2	Y_3
Y_1	1.00		
Y_2	.595	1.00	
Y_3	.448	.544	1.00

- Step 1: ab = .595
  ac = .448
  bc = .544
- Step 2: b = .595/a
  c = .488/a
  (.595/a)(.448/a) = .544
- Step 3: $.26656/a^2 = .544$
  a = .70
- Step 4: .70b = .595 b = .85
  .70c = .448 c = .64
- Step 5: $Var(e_1) = 1 - a^2 = .51$

Over-Identified Factor: 4+ Items

- Model is <u>over-identified</u> 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
0 factor means
4 loadings
4 item intercepts
4 error variances

OR

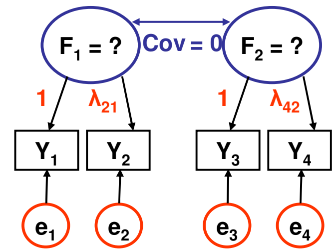
1 factor variance
1 factor mean
3 item loadings
3 item intercepts
4 error variances

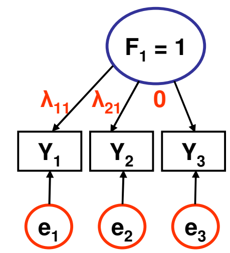
df = 14 – 12 = 2

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

Oops: Empirical Under-Identification

- Did your model blow up (errors instead of output)?
  - Make sure each factor is identified (scale set)
- Sometimes you can set up your model correctly and it will still blow up because of **empirical under-identification**
  - It's not you; **it's your data** – here are two examples





What to do with single indicators…

- Not all constructs need multiple items… Age? Sex?
  - In this case, you can define a latent factor for the construct if you wish, and fix the error variance of the single item to 0
- Some constructs need multiple items… but you may not have them (or only have a total score)
  - You can use known estimates of reliability to define a latent factor from the single item anyway
  - Fix factor loading = 1, error variance = Var(Y)(1-reliability)
    - Observed variance = 10 with reliability of .80? Error variance = 2
    - Error variance cannot be correlated with anything else then
  - If item reliability differs broadly across samples, may want to try alternative estimates to see how it impacts the model estimates

The Other Kind of Measurement Model…

The difference between principal components and factor analysis corresponds to different 'types' of items:

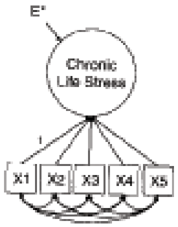
<u>Factor Model:</u>

- Composed of "Reflective" or "Effects" items
- Factor is **cause** of observed item responses
- Items **are** exchangeable and should be correlated
- **Is identified** with 3+ items (fit testable with 4+ items)

<u>Component Model:</u>

- Composed of "Formative" or "Emergent" or "Cause" items
- Component is **result** of observed item responses
- Items **are not** exchangeable and may not be correlated
- **Will not be identified** no matter how many items *without outside variables*

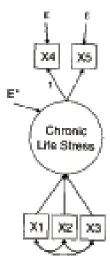
Formative (Component) Models
(see Brown p. 351-362)



Model Parameters:

4 factor loadings/regression paths
1 factor disturbance (variance left over)
10 item correlations
5 item variances

df = 15 – 20 = -5
Not identified



Model Parameters:

4 factor loadings/regression paths
1 factor disturbance (variance left over)
3 item correlations
5 item variances/errors

df = 15 – 13 = 2
Identified

What about Saving Factor Scores?

- Sometimes one might wish to save factor scores as generated in the factor model for use in other analyses… AVOID DOING THIS IF POSSIBLE.
- Although CFA factor scores are less indeterminate than EFA factor scores (because there is no rotation in CFA), they will still have issues:
  - They will be shrunken (i.e., pushed towards the mean, such that the observed variance of the factor scores will be less than the original factor variance).
  - They are just estimates of central tendency from a distribution for each person, not known values – and using estimates as known values in another model makes the relationships within that model look more precise than they are.
  - You CANNOT create factor scores by using the loadings as such:
    - $F = \lambda_{11}Y_1 + \lambda_{21}Y_2 + \lambda_{21}Y_3$… → This is a COMPONENT model, not a FACTOR model.
- It is most optimal to use a single-stage (SEM) estimation to obtain estimates of relationships among factors (i.e., estimate all factors at once).
- If sample size prohibits a single-stage analysis, a conservative (and effective) approach is to use "unit-weighting" → the Add Stuff Up model.

CTT vs. CFA

- What are the advantages of moving from CTT to CFA?
- More reasonable assumptions about items
  - CTT assumes tau-equivalence → equal factor loadings
  - CFA allows a test of whether each item relates to the factor, as well as whether different factor loadings across items are needed → would indicate some items are better than others
- Comparability across samples, groups, and time
  - CTT: No separation of observed item responses from true score
    - Sum across items = true score
  - CFA: Latent trait is estimated separately from item responses
    - Specific items may no longer matter; work in latent space

Wrapping Up…

- CFA is a framework in which observed items are predicted from latent factors (traits) and residual error
- Differs from EFA in matrix analyzed, how to get an interpretable solution, and how to assess model fit
- Goal is to reproduce observed covariance matrix of items with as few parameters as possible
  - Factor model makes very specific mathematical predictions about how observed items should relate to each other
  - Need at least 3 items per factor for the model to be identified; at least 4 items for model fit to be testable
- CFA framework offers significant advantages over CTT