# Confirmatory Factor Analysis (CFA) Part 1

- Topics:
  - Comparison of EFA and CFA
  - CFA model parameters
    - Measurement versus structural model parameters (parms)
    - How they combine to make patterns to match those in your data
  - Two parts of CFA model identification
    - Ensure each latent factor has a scale
    - Ensure model parameters are all estimable

# EFA vs. CFA: What gets analyzed

- **EFA: Correlation matrix (of items = indicators)**
  - Only correlations among observed item responses are used
  - Only a standardized solution is provided, so the original means and variances of the observed item responses are irrelevant

- **CFA: Covariance matrix (of items = indicators)**
  - Means, variances, and covariances of item responses are analyzed (where the latter two form the "item variance–covariance matrix")
    - "Item covariance matrix" is a shorter way of saying "item variance–covariance matrix"
  - Item response means historically have been ignored (but not by us!)
  - Output includes unstandardized AND standardized solutions
    - **Unstandardized** solution predicts the **item variance–covariance matrix** (regression solution retains original absolute information)
    - **Standardized** (STDYX) solution predicts the **item correlation matrix** (easier to interpret relative sizes of relationships as correlations)

# EFA vs. CFA: Interpretation

- **EFA: Rotation**
  - All items load on all **factors** (*aka*, latent traits), no matter what!
  - Goal is to pick a rotation that gives closest approximation to "simple structure" (clearly-defined factors, fewest cross-loadings)
  - No way of distinguishing latent variables due to "content" (traits being measured) from "method" (correlation created by common approach)
  - Summary: **It is NOT your data's job to tell you what they measure…**

- **CFA: Defining interpretation is your job in the first place!**
  - CFA is theory-driven: any structure becomes a testable hypothesis
  - You specify number of latent factors and their inter-correlations
  - You specify which items load on which latent factors (yes/no)
  - You specify any additional relationships for method/other covariance
  - You just need a clue; you don't have to be right (misfit is informative)

# EFA vs. CFA: Judging model fit

- **EFA: Eye-balls and Opinion**
  - #Factors? Scree-ish plots, interpretability…
  - Which rotation? Whichever makes the most sense… (to you): there is no way to say which is the "right" rotation (and interpretation)!
  - Which items load on each factor? Arbitrary cut-off of .3-.4ish
  - Standard errors found in newer software (but still infrequently used)

- **CFA: Inferential tests via Maximum Likelihood (ML or MLR)**
  - Global model fit test (and local model fit diagnostics)
  - Standard errors (and significance tests) for all parameters: item loadings, error variances, and error covariances
  - Ability to test appropriateness of model restrictions or model additions via tests for change in model fit

# EFA vs. CFA: Factor scores

- **EFA: Don't ever use factor scores from an EFA**
  - Factor scores are indeterminate (especially due to rotation)
  - Inconsistency in how factor models are applied to data
    - Factor model is based on common variance only (factors are predictors)
    - Summing items? That's using total variance (components are outcomes)

- **CFA: Factor scores *can* be used, but only if really necessary**
  - Best option: Test relations among latent factors directly through SEM
    - Factors can either be predictors ("exogenous" variables) or outcomes ("endogenous" variables) or both at once as needed (e.g., as mediators)
    - Relations between factors will be disattenuated for measurement error
  - Factor scores are less indeterminate in CFA, and could be used
    - In reality, though, factor scores are not known single values because they are modeled as *random* effects, not *fixed* effects per subject
    - Next-best option: Use "plausible values" or other two-stage approaches that acknowledge uncertainty in factor score estimates (stay tuned!)

# Confirmatory Factor Analysis (CFA)

- **In CFA, the unit of analysis is the ITEM (as in any LTMM):**

  $\boldsymbol{y_{is}} = \boldsymbol{\mu_i} + \boldsymbol{\lambda_i F_s} + \boldsymbol{e_{is}}$ → **both items AND subjects matter**

  Btw, in Brown (2015), my $\boldsymbol{F}$ = his $\boldsymbol{\xi}$ ("psi") and my $y_{is}$ = his $x_{is}$

  - Observed response for **item *i*** and **subject *s***
    = **intercept** of item *i* ($\boldsymbol{\mu}$)
    + **latent factor** of subject *s* ($\boldsymbol{F}$), weighted by item-specific **loading** $\boldsymbol{\lambda}$
    + **error** ($\boldsymbol{e}$) of item *i* and subject *s*

- **This looks like linear regression (with an unobserved $\boldsymbol{x}$)!**
  - $\boldsymbol{y_s} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_s} + \boldsymbol{e_s}$ → written for each item → $\boldsymbol{y_{is}} = \boldsymbol{\beta_{0i}} + \boldsymbol{\beta_{1i} x_s} + \boldsymbol{e_{is}}$
  - $\boldsymbol{\beta_{0i}}$ Intercept = $\boldsymbol{\mu_i}$ = expected item response $\boldsymbol{\hat{y}_{is}}$ **when** $\boldsymbol{F_s = 0}$
  - $\boldsymbol{\beta_{1i}}$ Slope of Factor = $\boldsymbol{\lambda_i}$ = expected difference in $\boldsymbol{y_{is}}$ for a one-unit difference in $\boldsymbol{F_s}$ ("difference" because these $\boldsymbol{F}$'s are for different people)
  - $\boldsymbol{e_{is}}$ Error (Residual) = how far off actual $\boldsymbol{y_{is}}$ is from predicted $\boldsymbol{\hat{y}_{is}}$
    - Residual has another meaning of "discrepancy" (for model misfit) in LTMMs, so I will try to use the term "error" for this item parameter instead

# Revisiting Vocabulary: Item Psychometric Properties

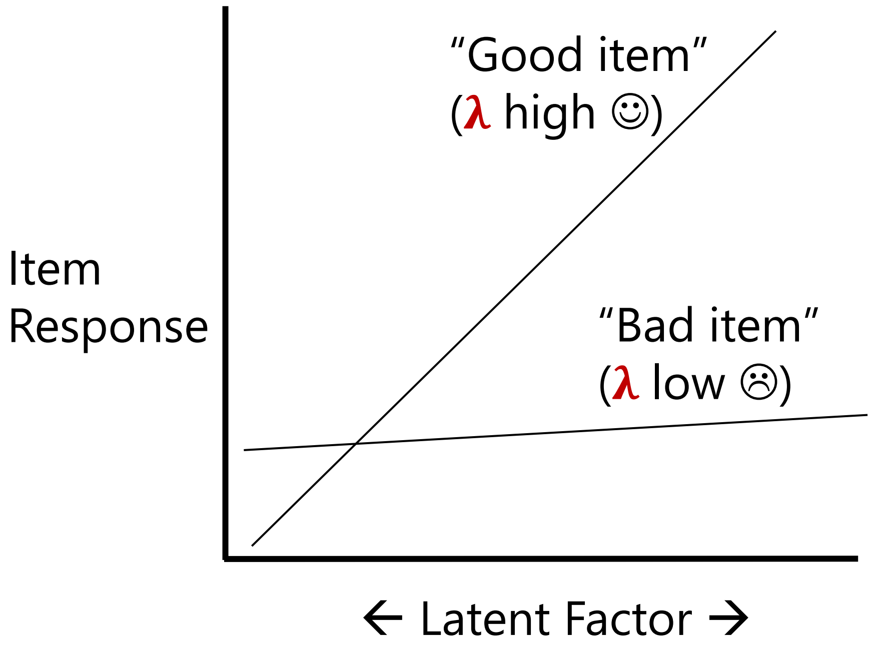
- **Item Discrimination**: How related each item is to the latent trait
  - In CTT, discrimination was given by the item–remainder correlation
    - The total score is the best estimate of the latent trait in CTT ($\boldsymbol{True_s} \rightarrow \boldsymbol{F_s}$ in CFA)
  - In **CFA**, discrimination is given by the **factor loading/slope ($\boldsymbol{\lambda_i}$)**
    - We now have a separate factor $\boldsymbol{F_s}$ (whose job is to recreate the item covariances)
    - **Stronger factor loadings indicate better, more discriminating items**
      - Use the "standardized" factor loadings when comparing across items (stay tuned!)

- **Item Difficulty/Severity**: Location of item on the latent trait metric
  - In CTT, difficulty is given by the item mean – which is sample-dependent
  - In **CFA**, difficulty is given by the **item intercept ($\boldsymbol{\mu_i}$)** – which is sample-free
  - In contrast to other latent trait models (such as IFA/IRT), difficulty (or "easiness" because we are using additive intercepts) are often ignored in CFA… here's why…

# Why Item Intercepts Are Often Ignored…

A **"good" item** has a **steep** slope (i.e., factor loading) in predicting the item response from the factor. Because this is a **linear slope**, the item is assumed to be equally discriminating (**equally good**) across the entire latent factor.

Similarly, a **"bad" item** has a **flatter** slope that is **equally bad** across the entire latent factor range (slope=0 means item is unrelated to factor).



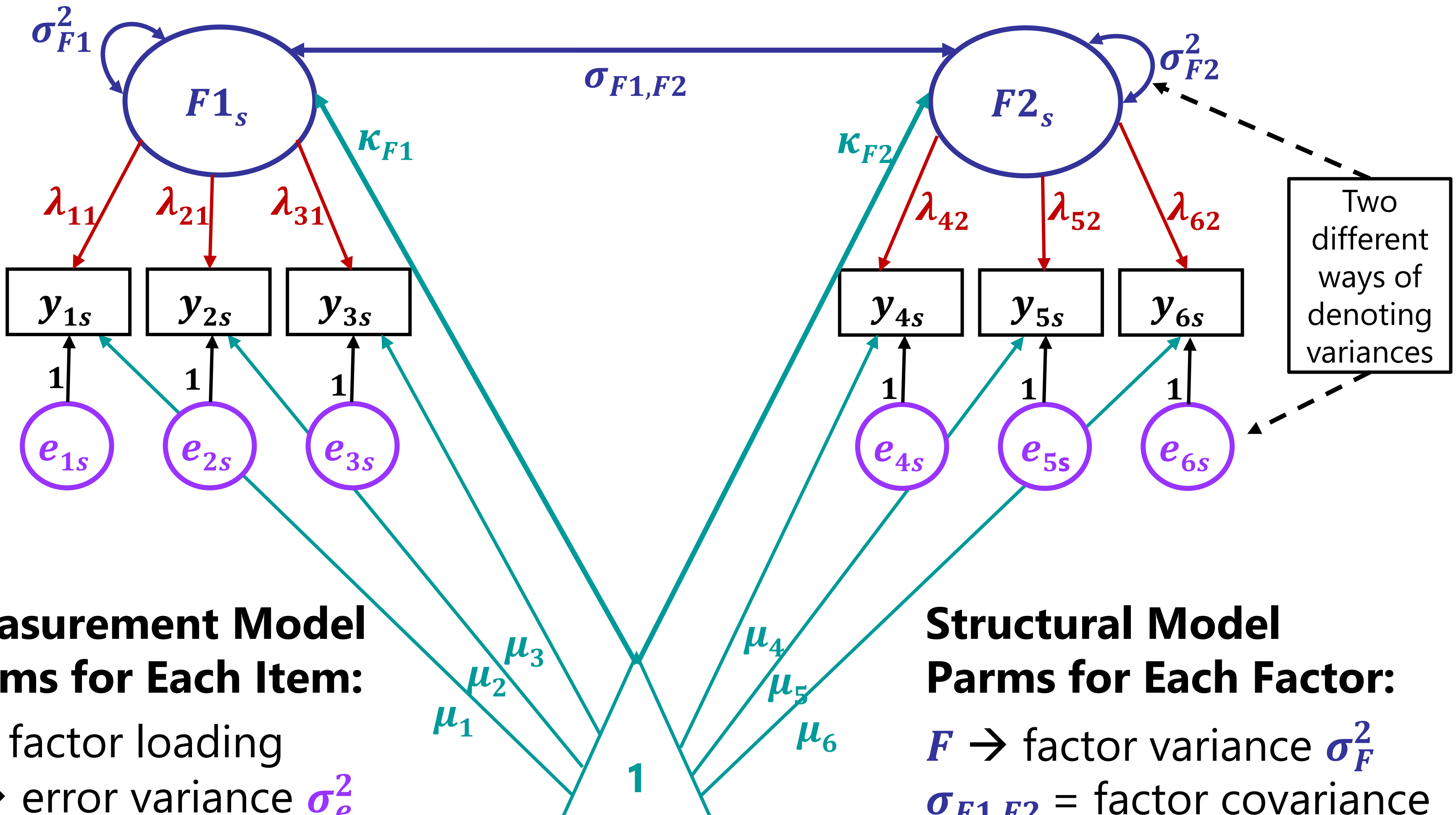

Here item intercepts are irrelevant in evaluating how "good" an item is, so they are not really needed in CFA.

But we will estimate item intercepts, because they are critical when:

- Testing factor mean differences in any latent factor model
- Items need to have a nonlinear slope in predicting the item response from the factor (IRT)

# Example Diagram of Two-Factor CFA Model

But some parameters will have to be **fixed** to known values for the model to be identified.




**Measurement Model Parms for Each Item:**

$\lambda$ = factor loading

$e$ → error variance $\sigma^2_e$

$\mu$ = intercept

**Structural Model Parms for Each Factor:**

$F$ → factor variance $\sigma^2_F$

$\sigma_{F1,F2}$ = factor covariance

$\kappa$ = factor mean

# Example "Congeneric" Two-Factor Model

- Measurement model for items ("indicators") 1–6 for subject *s*:
  - $y_{1s} = \mu_1 + \lambda_{11}F1_s + 0F2_s + e_{1s}$
  - $y_{2s} = \mu_2 + \lambda_{21}F1_s + 0F2_s + e_{2s}$
  - $y_{3s} = \mu_3 + \lambda_{31}F1_s + 0F2_s + e_{3s}$
  - $y_{4s} = \mu_4 + 0F1_s + \lambda_{42}F2_s + e_{4s}$
  - $y_{5s} = \mu_5 + 0F1_s + \lambda_{52}F2_s + e_{5s}$
  - $y_{6s} = \mu_6 + 0F1_s + \lambda_{62}F2_s + e_{6s}$

You decide **how many factors** and which items they predict ("congeneric" → diff item parms)

Unstandardized loadings ($\lambda_i$) are **linear slopes** predicting the item response ($y_{is}$) from the factors ($F_s$). **Thus, the model assumes a linear relationship between each factor and item response.**

Intercepts ($\mu_i$) are the expected item responses ($\hat{y}_{is}$) when *all factors predicting that item* = 0.

Here is a more general matrix equation for these 6 item-specific equations:

$$\mathbf{Y} = \mathbf{M} + \mathbf{\Lambda U} + \mathbf{E}$$

where $\mathbf{Y}$, $\mathbf{M}$, and $\mathbf{E}$ = 6 × 1 matrices (because each item gets one value of each); $\mathbf{\Lambda}$ = 6 × 2 matrix and $\mathbf{U}$ = 2 × 1 matrix (because two factors)

$\mathbf{M}$ = capital $\mu$, $\mathbf{\Lambda}$ = capital $\lambda$, and $\mathbf{U}$ holds the $F$ values (because $F$ is already capitalized and because random effects = factor scores)

# What the Model is *Really* Doing

- **We don't have observed *F* values** (called "factor scores") by which to predict the item responses, but **we don't need them**!
  - The ***F* values are random effects** assumed to have a multivariate normal (MVN) distribution, which means all we need to estimate the model are parameters for their means, variances, and covariances

- **The role of the CFA model parameters is to recreate the summary statistics of the item (indicator) responses**
  - In CFA, these are the item means, variances, and covariances (Pearson-style), which are sufficient statistics assuming a multivariate normal conditional distribution for the item responses from the same subject
  - In IRT and IFA, item summary statistics will vary by model and estimator (because variance is not separately estimated for categorical outcomes)

- How well the **observed** item summary statistics **match** those **recreated by the CFA model** is the primary basis of model fit…
  - Just as it is in longitudinal models with balanced time! (e.g., AR1 vs UN)

# How CFA Models Recreate Item Means

Matrix equation: $\bar{\mathbf{Y}} = \mathbf{M} + \boldsymbol{\Lambda}\mathbf{K}$

$\bar{\mathbf{Y}}$ = Recreated item mean vector built from matrices holding:

**$\mathbf{M}$ = ("Mu") item intercepts ($\mu$)**

**$\boldsymbol{\Lambda}$ = ("Lambda") item factor loadings ($\lambda$)**

**$\mathbf{K}$ = ("Kappa") factor means ($\kappa$)**

Btw, the factor means are usually fixed to 0 for identification (stay tuned!), such that the item intercepts then take on the values of the item means directly.

But because that's not always the case, **we will interpret $\mu$ as an item intercept specifically → item response when all its factor score predictors = 0**

**Scalar version of CFA model:**

$$y_{is} = \mu_i + \lambda_{i1}F1_s + \lambda_{i2}F2_s + e_{is}$$

Below, for our example model for 6 items and 2 congeneric factors (one factor per item):

$$\bar{y}_1 = \mu_1 + \lambda_{11}\kappa_{F1} + 0\kappa_{F2}$$
$$\bar{y}_2 = \mu_2 + \lambda_{21}\kappa_{F1} + 0\kappa_{F2}$$
$$\bar{y}_3 = \mu_3 + \lambda_{31}\kappa_{F1} + 0\kappa_{F2}$$
$$\bar{y}_4 = \mu_4 + 0\kappa_{F1} + \lambda_{42}\kappa_{F2}$$
$$\bar{y}_5 = \mu_5 + 0\kappa_{F1} + \lambda_{52}\kappa_{F2}$$
$$\bar{y}_6 = \mu_6 + 0\kappa_{F1} + \lambda_{62}\kappa_{F2}$$

# Recreated Item Variance–Covariance Matrix

Matrix equation: $\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Phi\Lambda}^T + \boldsymbol{\Psi}$

Btw, more commonly, my $\sigma_F^2$ **is** $\phi$ **("phi")** and my $\sigma_e^2$ **is** $\psi$ **("psi")**

**Σ** = ("Sigma") is model-recreated item variance–covariance matrix built from:

- **Λ = ("Lambda") item factor loadings**
- **Φ = ("Phi") factor variances and covariances**
- **$\Lambda^T$ = item factor loadings transposed (→$\lambda^2$)**
- **Ψ = ("Psi") item error (residual) variances)**

$$
\begin{pmatrix}
\sigma^2_{y1} & \sigma_{y1,y2} & \sigma_{y1,y3} & \sigma_{y1,y4} & \sigma_{y1,y5} & \sigma_{y1,y6} \\
\sigma_{y2,y1} & \sigma^2_{y2} & \sigma_{y2,y3} & \sigma_{y2,y4} & \sigma_{y2,y5} & \sigma_{y2,y6} \\
\sigma_{y3,y1} & \sigma_{y3,y2} & \sigma^2_{y3} & \sigma_{y3,y4} & \sigma_{y3,y5} & \sigma_{y3,y6} \\
\sigma_{y4,y1} & \sigma_{y4,y2} & \sigma_{y4,y3} & \sigma^2_{y4} & \sigma_{y4,y5} & \sigma_{y4,y6} \\
\sigma_{y5,y1} & \sigma_{y5,y2} & \sigma_{y5,y3} & \sigma_{y5,y4} & \sigma^2_{y5} & \sigma_{y5,y6} \\
\sigma_{y6,y1} & \sigma_{y6,y2} & \sigma_{y6,y3} & \sigma_{y6,y4} & \sigma_{y6,y5} & \sigma^2_{y6}
\end{pmatrix}
$$

$$
= \begin{pmatrix}
\lambda_{11} & 0 \\
\lambda_{21} & 0 \\
\lambda_{31} & 0 \\
0 & \lambda_{42} \\
0 & \lambda_{52} \\
0 & \lambda_{62}
\end{pmatrix}
\begin{pmatrix}
\sigma^2_{F1} & \sigma_{F1,F2} \\
\sigma_{F2,F1} & \sigma^2_{F2}
\end{pmatrix}
\begin{pmatrix}
\lambda_{11} & \lambda_{21} & \lambda_{31} & 0 & 0 & 0 \\
0 & 0 & 0 & \lambda_{42} & \lambda_{52} & \lambda_{62}
\end{pmatrix}
$$

$$
+ \begin{pmatrix}
\sigma^2_{e1} & 0 & 0 & 0 & 0 & 0 \\
0 & \sigma^2_{e2} & 0 & 0 & 0 & 0 \\
0 & 0 & \sigma^2_{e3} & 0 & 0 & 0 \\
0 & 0 & 0 & \sigma^2_{e4} & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma^2_{e5} & 0 \\
0 & 0 & 0 & 0 & 0 & \sigma^2_{e6}
\end{pmatrix}
$$

# Recreated Item Variance–Covariance Matrix

## $\Sigma = \Lambda\Phi\Lambda^T + \Psi$ → Model-Implied Σ Matrix:

*Sigma for model = Lambda*(Phi)*Lambda$^T$ + Psi*

The **loadings** control how related items from the same factor are predicted to be

**Items within Factor 1**

$$\begin{pmatrix}
\lambda_{11}^2\sigma_{F1}^2+\sigma_{e1}^2 & \lambda_{11}\sigma_{F1}^2\lambda_{21} & \lambda_{11}\sigma_{F1}^2\lambda_{31} & \lambda_{11}\sigma_{F2,F1}\lambda_{42} & \lambda_{11}\sigma_{F2,F1}\lambda_{52} & \lambda_{11}\sigma_{F2,F1}\lambda_{62} \\
\lambda_{21}\sigma_{F1}^2\lambda_{11} & \lambda_{21}^2\sigma_{F1}^2+\sigma_{e2}^2 & \lambda_{21}\sigma_{F1}^2\lambda_{31} & \lambda_{21}\sigma_{F2,F1}\lambda_{42} & \lambda_{21}\sigma_{F2,F1}\lambda_{52} & \lambda_{21}\sigma_{F2,F1}\lambda_{62} \\
\lambda_{31}\sigma_{F1}^2\lambda_{11} & \lambda_{31}\sigma_{F1}^2\lambda_{21} & \lambda_{31}^2\sigma_{F1}^2+\sigma_{e3}^2 & \lambda_{31}\sigma_{F2,F1}\lambda_{42} & \lambda_{31}\sigma_{F2,F1}\lambda_{52} & \lambda_{31}\sigma_{F2,F1}\lambda_{62} \\
\lambda_{42}\sigma_{F2,F1}\lambda_{11} & \lambda_{42}\sigma_{F2,F1}\lambda_{21} & \lambda_{42}\sigma_{F2,F1}\lambda_{31} & \lambda_{42}^2\sigma_{F2}^2+\sigma_{e4}^2 & \lambda_{42}\sigma_{F2}^2\lambda_{52} & \lambda_{42}\sigma_{F2}^2\lambda_{62} \\
\lambda_{52}\sigma_{F2,F1}\lambda_{11} & \lambda_{52}\sigma_{F2,F1}\lambda_{21} & \lambda_{52}\sigma_{F2,F1}\lambda_{31} & \lambda_{52}\sigma_{F2}^2\lambda_{42} & \lambda_{52}^2\sigma_{F2}^2+\sigma_{e5}^2 & \lambda_{52}\sigma_{F2}^2\lambda_{62} \\
\lambda_{62}\sigma_{F2,F1}\lambda_{11} & \lambda_{62}\sigma_{F2,F1}\lambda_{21} & \lambda_{62}\sigma_{F2,F1}\lambda_{31} & \lambda_{62}\sigma_{F2}^2\lambda_{42} & \lambda_{62}\sigma_{F2}^2\lambda_{52} & \lambda_{62}^2\sigma_{F2}^2+\sigma_{e6}^2
\end{pmatrix}$$

**Items within Factor 2**

The only reason why items from different factors should be related is the **covariance** between the two factors

The **loadings** also control how much of the item response is due to factor versus error

# The **Role** of the CFA Model Parameters

- Data going in to be recreated by the CFA model parameters = item mean vector + item variance–covariance matrix

- CFA **item intercepts** ($\mu_i$) try to recreate the **item means**
  - Item *means* are unconditional; item *intercepts* are conditional on $F_s = 0$
  - When each item gets its own intercept (the usual case), the item means *should** be perfectly recreated (**so no room for mis-fit or mis-prediction**)

- CFA **item error variances** ($\sigma^2_{e_i}$) try to recreate the **item variances**
  - Item variances are unconditional; item error variances are conditional (= leftover variance after accounting for the contribution of the factor)
  - When each item gets its own error variance (usual case), the item variances *should** be perfectly recreated (**so no room for misfit or mis-prediction**)

- CFA **item factor loadings** ($\lambda_i$) try to recreate the **item covariances**
  - Given 3+ items, there will be more covariances among items to predict than item factor loadings to recreate them, **thus creating room for misfit**

# 2 Types of CFA Parameter Solutions

- Unstandardized → predicts scale-sensitive original item response:
  - **Uses scale-specific regression model:** $y_{is} = \mu_i + \lambda_i F_s + e_{is}$
  - ***Useful when comparing solutions across samples (when original values matter)***
  - Unstandardized model parameters predict ***item means and variance–covariance matrix***
  - Note the solution asymmetry: item parameters $\mu_i$ and $\lambda_i$ are given in the original metric, but $e_{is}$ is replaced as the error variance across subjects for that item (squared metric)
  - $Var(y_i) = [\lambda_i^2 * Var(F_s)] + Var(e_{is})$ → $\sigma_{y_i}^2 = [\lambda_i^2 * \sigma_F^2] + \sigma_{e_i}^2$

In Mplus, use STDYX

- Standardized → Rescaled so $M(y_i) = 0$; $Var(y_i) = Var(F_s) = 1$:
  - ***Useful when comparing items within a model → relative values on same scale***
  - Standardized model predicts the **item correlation matrix (b/c item means = 0)**
  - Standardized intercept = $\mu_i$ / $SD(y_{is})$ → not typically reported (and mostly unhelpful)
  - Standardized factor loading = $[\lambda_i * SD(F_s)]$ / $SD(y_{is})$ = **item correlation with factor**
  - Standardized factor loading$^2$ = item $R^2$ = proportion of item variance DUE to the factor
  - Standardized error variance = $1 - stand\ \lambda_i^2$ = prop. item variance due to *NOT* the factor

# CFA Model Predictions: $F1$ → $y_{1s}$–$y_{3s}$, $F2$ → $y_{4s}$–$y_{6s}$

Items from same factor (room for misfit=bad recreation):

- Unstandardized solution: Covariance$(y_1, y_3)$ = $\lambda_{11} * Var(F1) * \lambda_{31}$
- Standardized solution: Correlation$(y_1, y_3)$ = $\lambda_{11} * (1) * \lambda_{31}$ ← **std loadings**
- ONLY reason for correlation is their common factor (local independence, LI)

Items from different factors (room for misfit=bad recreation):

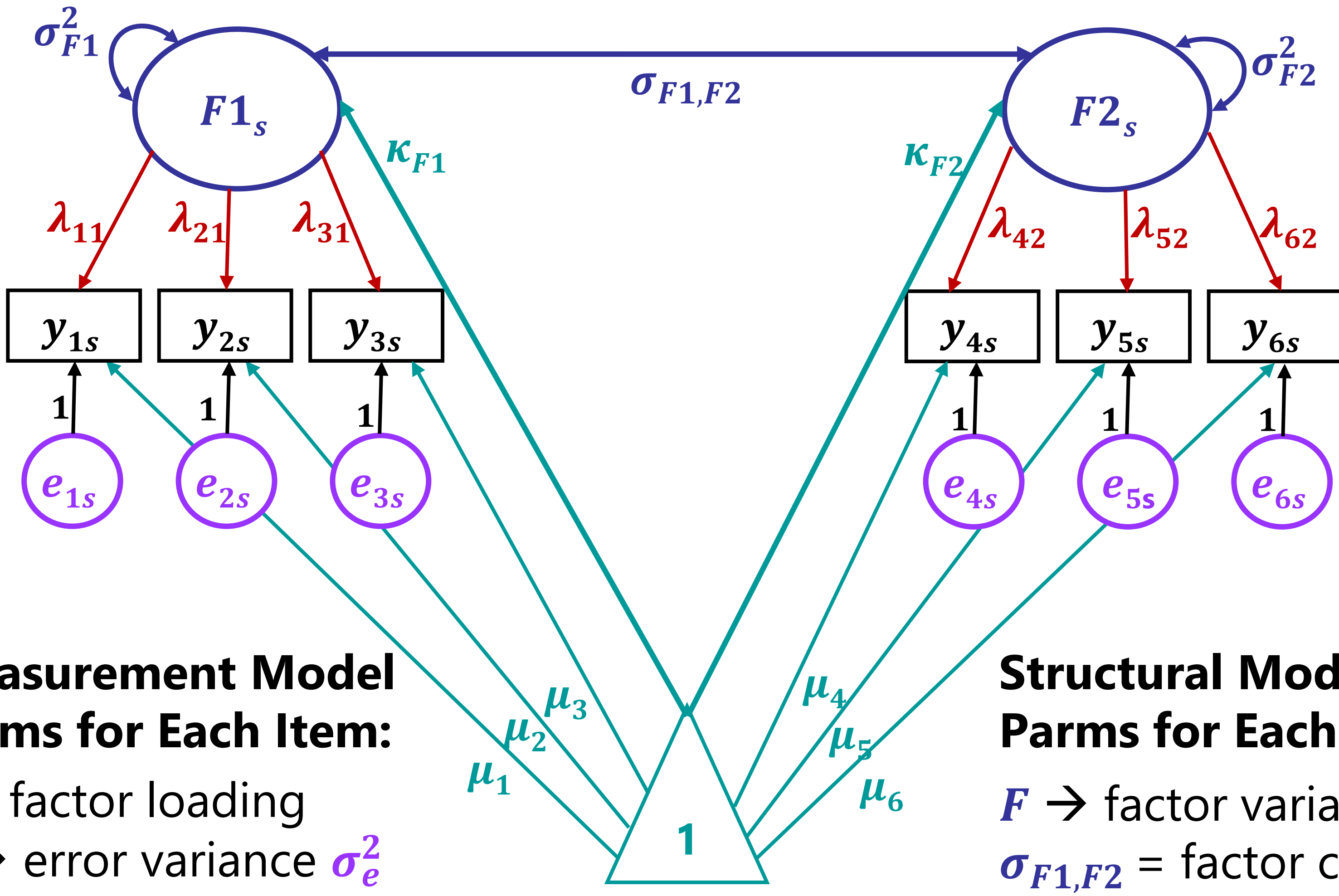
- Unstandardized: Covariance$(y_1, y_6)$ = $\lambda_{11} * Cov(F1, F2) * \lambda_{62}$
- Standardized: Correlation$(y_1, y_6)$ = $\lambda_{11} * Cor(F1, F2) * \lambda_{62}$ ← **std loadings**
- ONLY reason for correlation is the correlation between factors (again, LI)

Variances are additive (and should be reproduced correctly):

- $Var(y_1) = (\lambda_{11}{}^2) * Var(F1) + Var(e_i)$ → **note imbalance of $\lambda^2$ and $e_i$** ($e_i$ is not a model parameter – its variance across subjects is instead)
- Same equation applies to both unstandardized and standardized solutions

# Example Diagram of Two-Factor CFA Model

But some parameters will have to be **fixed** to known values for the model to be identified.




**Measurement Model Parms for Each Item:**

$\lambda$ = factor loading

$e$ → error variance $\sigma^2_e$

$\mu$ = intercept

**Structural Model Parms for Each Factor:**

$F$ → factor variance $\sigma^2_F$

$\sigma_{F1,F2}$ = factor covariance

$\kappa$ = factor mean

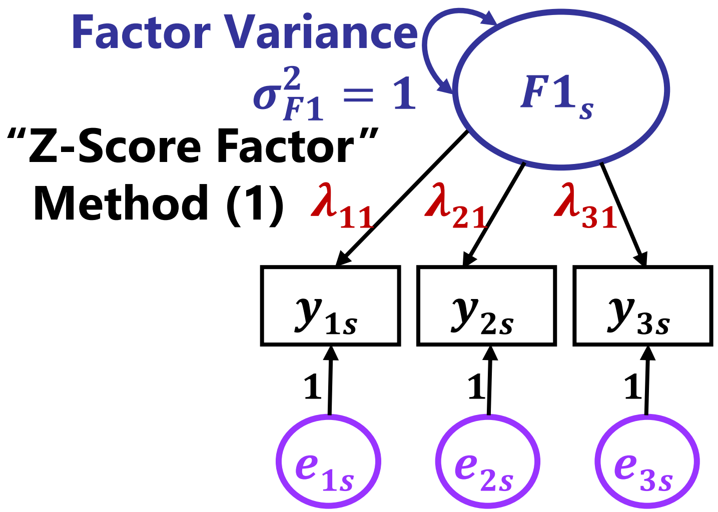
# Two Parts of Model Identification

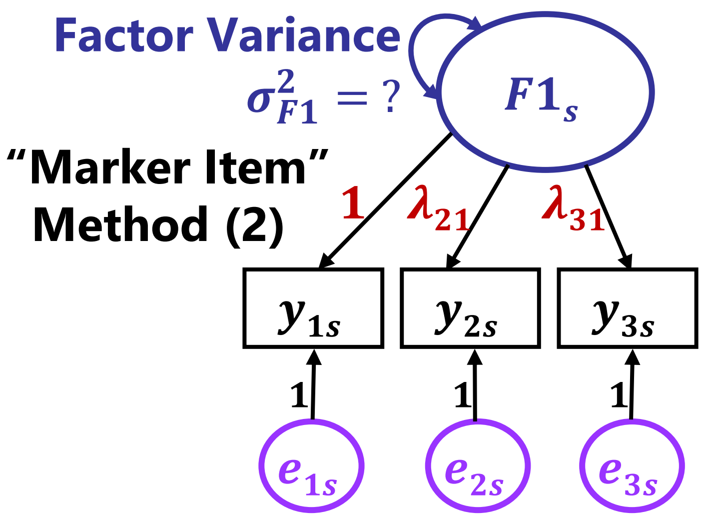
- **Part 1: Each latent factor has a scale**
  - Each latent factor needs a mean and a variance
  - Necessary but not sufficient for estimating the CFA model

- **Part 2: Ensure the CFA model is estimable**
  - Data going in versus estimated parameters going out:
    - Item **means** → item **intercepts** (usually **1:1** ratio)
    - Item **variances** → item **error variances** (usually **1:1** ratio)
    - Item **covariances** → item **factor loadings** (must have **ratio ≥ 1**)
  - In practice, this means the number of estimated loadings may not exceed the number of observed item covariances
  - This also means that each pair of variables has only one direct relationship (more of a concern in path models)

# CFA Model Identification Part 1:

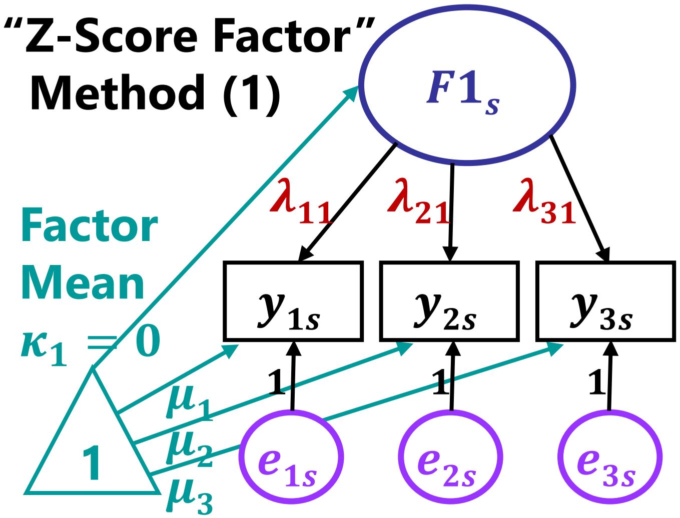
## *Create a Scale for the Latent Factor Variance*

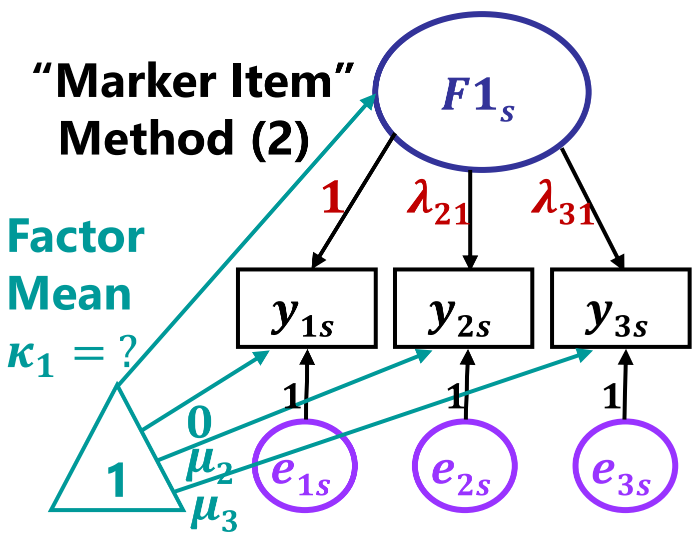



- The factor doesn't exist, so **it needs a scale** (it needs a mean and variance):
- There are two **equivalent** options to create a scale for the factor **VARIANCE**:
  - **(1) Fix factor variance to 1: "z-score"**
    - Factor SD is same as standard z-score (but any constant>0 would also be ok)
    - **Shouldn't** be used in models with higher-order factors (coming later in this course)
  - **(2) Fix a "marker item" loading to 1**
    - Factor variance is then estimated the "reliable" part of the marker item variance
    - e.g., Std. loading = 0.9, item variance = 16? Factor variance = $(0.9^2)$*16 = 12.96
    - Can cause the model to blow up if marker item has no correlation with the factor at all

# CFA Model Identification Part I:
## *Create a Scale for the Latent Factor <u>Mean</u>*



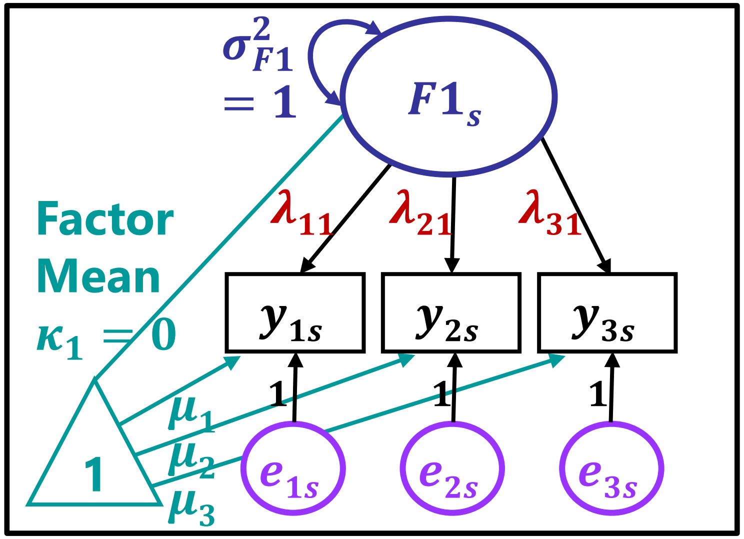



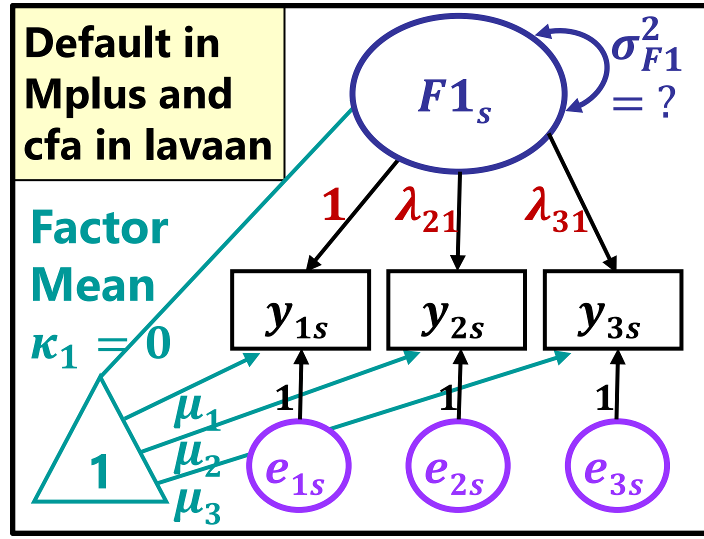

- The factor doesn't exist, so **it needs a scale** (it needs a mean and variance):
- There are two **equivalent** options to create a scale for the factor **MEAN**:
  - **(1) Fix factor mean to 0: "z-score"**
    - Factor mean is same as standard z-score (but any constant would also be ok)
    - **Can** be used in models with higher-order factors (coming later in the course)
    - Item intercepts = item means in this case
  - **(2) Fix a "marker item" intercept to 0**
    - Factor mean = mean of marker item
    - Item intercepts = expected item responses when factor = 0 (→ marker = 0, which may not make sense for all item response scales), not often used (except growth models)

# Part 1: Possible Factor Means and Variances

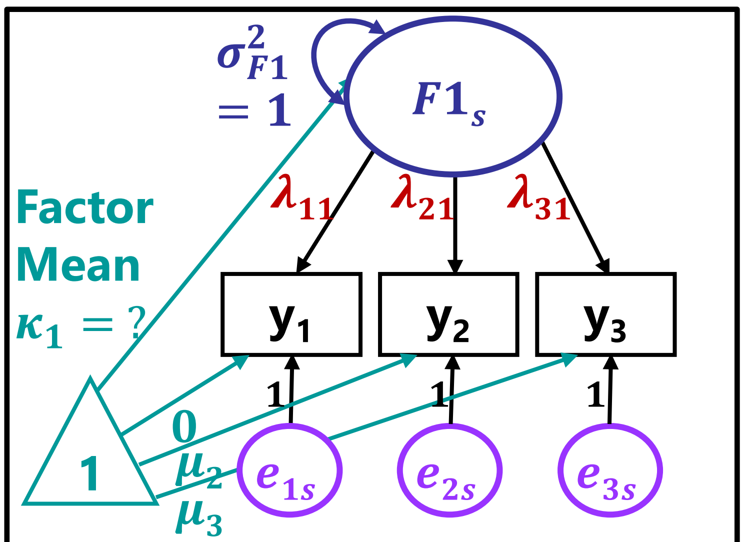
**Factor Variance = 1 (fixed)** | **Factor Variance Estimated**

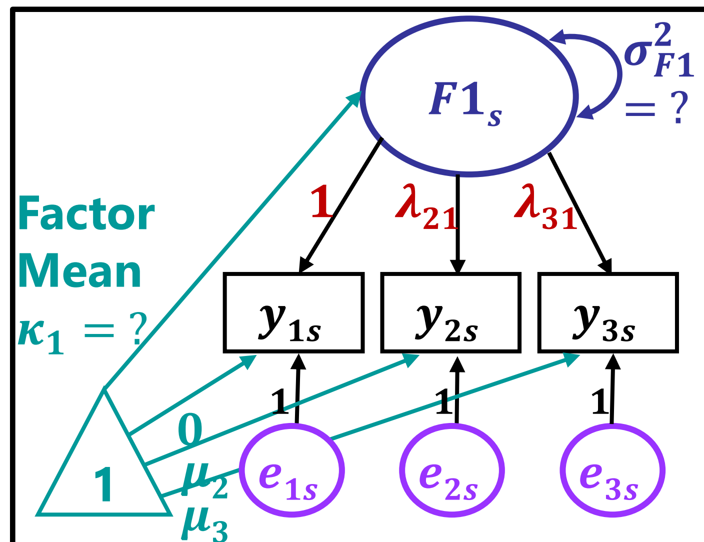
**Factor Mean $\kappa_1 = 0$ (fixed)**




**Factor Mean $\kappa_1 = ?$ (free)**



# Part 2 of CFA Model Identification

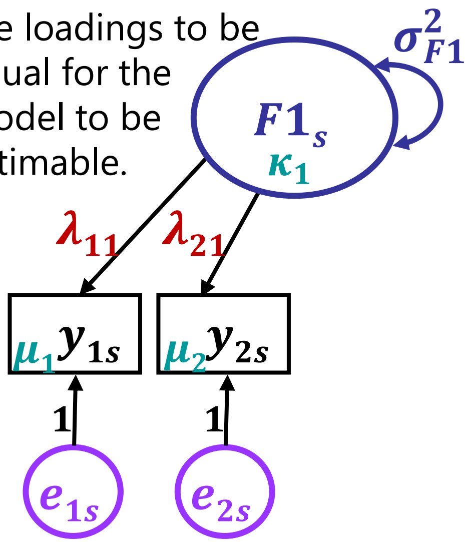
- **After scaling the factors**, then try to reproduce observed item means and variance–covariance matrix using as few estimated parameters as possible
  - (Robust) Maximum likelihood used to estimate model parameters
    - **Measurement Model**: Item intercepts, item factor loadings, item error variances
    - **Structural Model**: Factor means, factor variances, factor covariances
  - Global model fit = difference between model-recreated and data-observed variance–covariance matrix (but only covariances usually contribute to misfit)

- How many possible parameters can you estimate: What is the total $\mathrm{DF}$?
  - **Total $\mathbf{DF}$ =** $\boldsymbol{\frac{v(v+1)}{2} + v}$ where $\boldsymbol{v}$ **is the # items** (NOT people, like usual)
    - Total $\mathrm{DF}$ = number of item means, variances, and covariances
    - e.g., if $v = 4$ items, then Total $\mathrm{DF} = \frac{4(4+1)}{2} + 4 = 14$
  - **Model $\mathbf{DF}$** = data input − model output
  - **Model $\mathbf{DF}$** = # possible parameters − # estimated parameters
  - **Model $\mathbf{DF}$** = # parameter left over (could have been estimated, but didn't)

# Under-Identified Factor: 2 Items

- Model is **under-identified** if there are more unknown parameters then item variances, covariances, and means with which to estimate them
  - Model **cannot be estimated** because there are an infinite number of different parameter estimates that would result in the same (and perfect) fit
  - Example: x + y = 7 ??

You'd have to constrain the loadings to be equal for the model to be estimable.




In other words, the assumptions required to calculate two-score reliability in CTT are the result of model under-identification.

Total possible **DF** = unique pieces of data = **5**

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

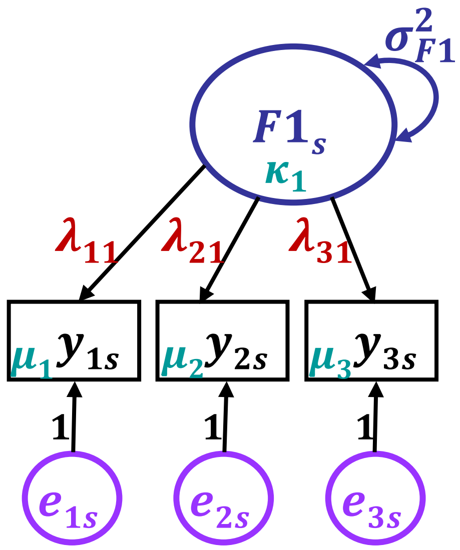
**Model DF = 5 – 6 = – 1**

If $cor(y_1, y_2) = .64$, then:

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

# Just-Identified Factor: 3 Items

- Model is **just-identified** if there are **as many** unknown parameters as item variances, covariances, and means with which to estimate them
  - The model is **estimable**, so the parameter estimates have a unique solution
  - But parameters will **perfectly reproduce** the observed variance–covariance matrix, so **model fit is not testable**—it's just a re-arrangement of the data
  - Example: Solve x + y = 7, 3x – y = 1




Total possible **DF** = unique pieces of data = **9**

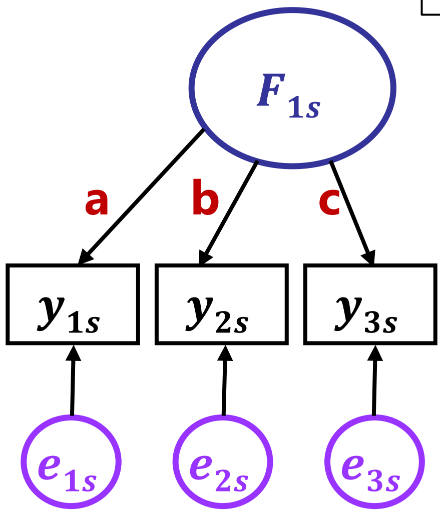
| | | |
|---|---|---|
| **0 factor variances** | | **1 factor variance** |
| **0 factor means** | | **1 factor mean** |
| **3 item loadings** | **OR** | **2 item loadings** |
| **3 item intercepts** | | **2 item intercepts** |
| **3 error variances** | | **3 error variances** |

**Model DF = 9 – 9 = 0**

Not really a model—more like a description

# Example: Solving a Just-Identified Model

Using the standardized solution, in which all item variances = 1 and factor variance = 1; item correlations shown below

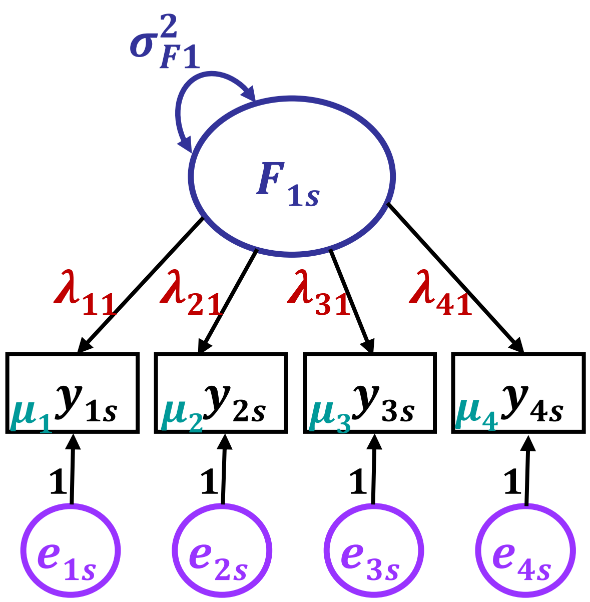



| | $y_1$ | $y_2$ | $y_3$ |
|---|---|---|---|
| $y_1$ | 1.00 | | |
| $y_2$ | .595 | 1.00 | |
| $y_3$ | .448 | .544 | 1.00 |

- 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: $.70b = .595 \quad b = .85$
  $.70c = .448 \quad c = .64$
- Step 5: $Var(e_1) = 1 - a^2 = .51$

# Over-Identified Factor: 4+ Items

- Model is **over-identified** if there are fewer unknown parameters than item variances, covariances, and means with which to estimate them
  - The model is **estimable**, so the parameter estimates have a unique solution
  - But now the parameters will NOT perfectly reproduce the observed matrix **→ if Model DF > 0, we can now test model fit!**




Total possible **DF** = unique pieces of data = **14**

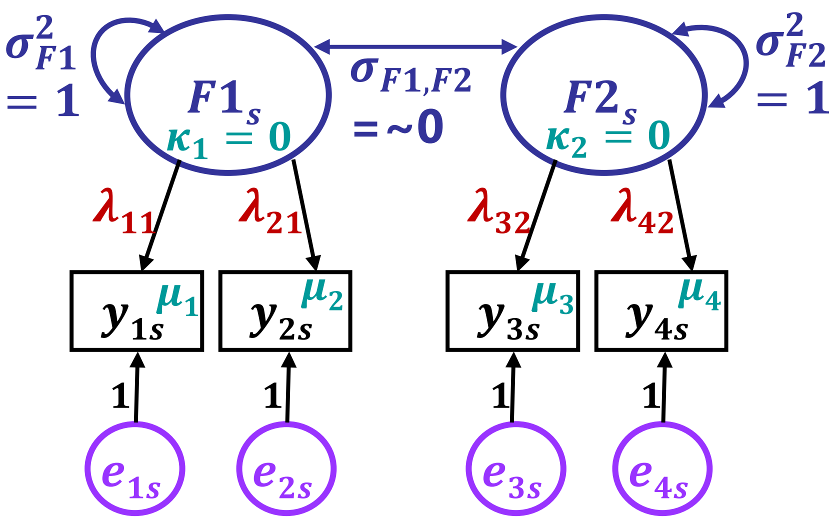
| | | |
|---|---|---|
| **0 factor variances** | | **1 factor variance** |
| **0 factor means** | | **1 factor mean** |
| **4 item loadings** | **OR** | **3 item loadings** |
| **4 item intercepts** | | **3 item intercepts** |
| **4 error variances** | | **4 error variances** |

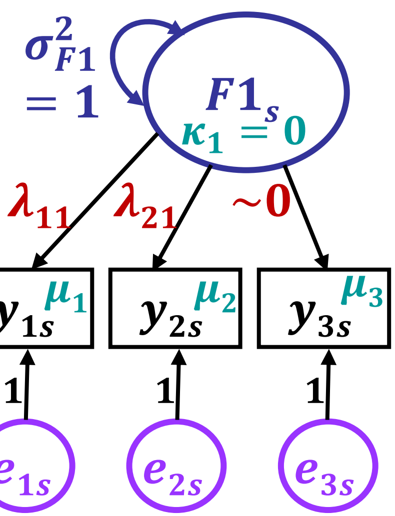
**Model DF = 14 – 12 = 2**

**Model fit:** Did we do a "good enough" job reproducing the item covariance matrix with 2 fewer parameters than it was possible to use?

# Oops: Empirical Under-Identification

- Did your model blow up (errors instead of output)? Double-check:
  - Part 1: Make sure each factor has a scale: a mean and a variance
  - Part 2: Make sure you aren't estimating more parameters than you have DF
- 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 of when these models should have been identified, but weren't because of an unexpected 0 relationship



# That Other Kind of Measurement Model… *(i.e., as used in principal components or partial least squares alternatives to CFA)*

Here is a difference between principal components and factor analysis in terms of "types" of items…

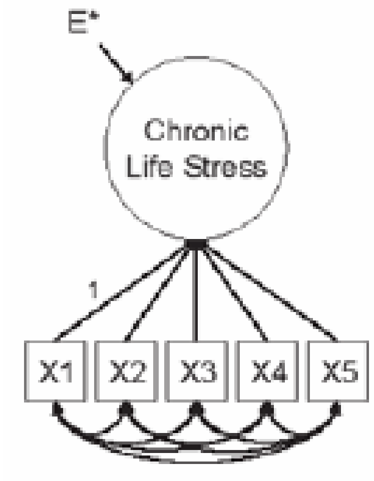
**Factor Model (CFA):**

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

**Component Model:**

- Composed of "Formative" or "Emergent" or "Cause" items
- Component is **result** of observed item responses
- Items may not be correlated
- **Will not be identified** no matter how many items *without additional variables in the model*

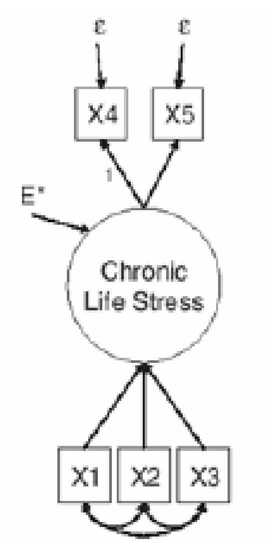
# Formative (Component) Models
## (see Brown 2015 ch. 8 pp. 322-331)




**Model A Parameters:**
4 factor loadings/regression paths
1 factor disturbance (variance left over)
10 item correlations
5 item variances
5 item means
$DF = 20 - 25 = -5$

**Not identified**

Formative measurement models are not identified without including other outcomes or predictors of the formative latent factor




**Model C Parameters:**
4 factor loadings/regression paths
1 factor disturbance (variance left over)
3 item correlations
5 item variances/error variances
5 item means/intercepts
$DF = 20 - 18 = 2$

**Identified**

Model C has both formative and reflective indicators—the latter might also be "outcomes"

# Summary of CFA Part 1

- **CFA is a linear model** in which continuous observed item responses are predicted from continuous latent factors and error
  - Goal: recreate observed **item means and variance–covariance matrix** using parameters in measurement model (item intercepts, loadings, error variances) and in structural model (factor means, variances, covariances)
  - Factor model makes specific testable mathematical predictions about how item responses should relate to each other: **loadings predict covariances**
  - Need at least 3 items per latent factor for the model to be identified; need **at least 4 items per latent factor for model fit to be testable**

- CFA framework offers significant advantages over CTT by offering the potential for comparability across samples, groups, and time
  - CTT: No separation of observed item responses from true score
    - Sum across items = true score; item properties belong to that sample only
  - CFA: Latent factor is estimated *separately* from item responses
    - Separates interpretation of person trait levels from specific items given
    - Separates interpretation of item properties from specific persons in sample