Measurement Invariance (MI) in CFA and Differential Item Functioning (DIF) in IRT/IFA

- Topics:
  - What are MI and DIF?
  - Testing measurement invariance in CFA
  - Testing differential item functioning in IRT/IFA

  - Btw: If you have many groups to compare, a newer "alignment" method may also be useful (see Asparouhov & Muthén, 2014)
  - Btw: If you don't have groups to compare, see slide 34
  - Slides 35 and 36 added to elaborate on configural invariance

The Big Picture

- **In CFA**, we can assess "**measurement invariance**" (**MI**), also known as "factorial invariance" or "measurement equivalence"
- Concerns the extent to which the psychometric properties of the observed indicators are transportable or generalizable across groups (e.g., gender, country) or over time/conditions
  - In other words, we are testing whether the indicators measure the same construct *in the same way* in different groups or over time/conditions
  - If so, then indicator responses should depend only on latent trait scores, and not on group membership or time/conditions, such that observed response differences are only caused by TRUE differences in the trait
- **In IRT/IFA**, lack of measurement invariance is known as "**differential item functioning**" (**DIF**), but it's the same idea
  - But note the inversion: Measurement Invariance = *Non*-DIF
Measurement *Non*-Invariance = DIF

2 Distinct Types of Invariance

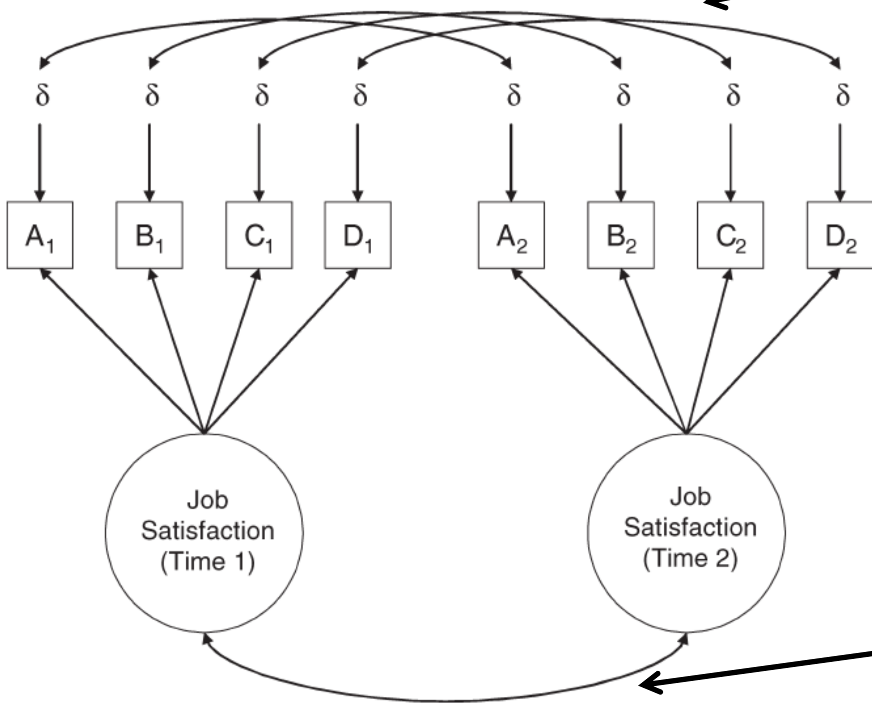
- **Measurement Invariance** concerns how the indicators (items) measure the latent trait across groups or time/conditions
  - An invariant measurement model has the same factor **loadings**, item **intercepts/thresholds**, and **residual variances** (and covariances)
  - Measurement model invariance is a precursor to ANY group or time/condition comparison (whether it is explicitly tested or not)
  - **It's NOT ok** if you don't have at least partial measurement invariance to make subsequent comparisons across groups or time/condition

- **Structural Invariance** concerns how the latent traits are distributed and related across groups or time/condition
  - An invariant structural model has the same **factor variances, factor covariances** (or same higher-order structure), and **factor means**
  - Given (at least partial) measurement invariance, **it IS ok** if you don't have structural invariance, **because trait differences may be real**

Model Options for Testing Invariance

- Invariance testing in CFA (or DIF testing in IRT/IFA) proceeds differently depending on the type of groups to be compared
- **Independent groups?** Use a "**multiple-group**" model
  - Estimate separate group-specific factor models **simultaneously**
  - Use GROUP= or KNOWNCLASS= in Mplus and separate MODEL: per group
  - An alternative approach of MIMIC models, in which the grouping variable is entered as a predictor, does not allow testing of equality of factor loadings or factor variances (so MIMIC is less useful than a full multiple-group model)
- **Dependent** (longitudinal, repeated, dyadic) groups?
  - All indicator responses go into **SAME model**, with separate factors per occasion/condition (allowing all factor covariances by default)
  - Usually, the same indicators also have residual covariances by default (which may have to be implemented as method/specific factors in IRT/IFA)
  - Given measurement invariance, growth modeling of the latent traits can serve as a specific type of structural invariance testing
  - It is INCORRECT to use a multiple-group model if the groups are dependent

Repeated Measures Invariance Model



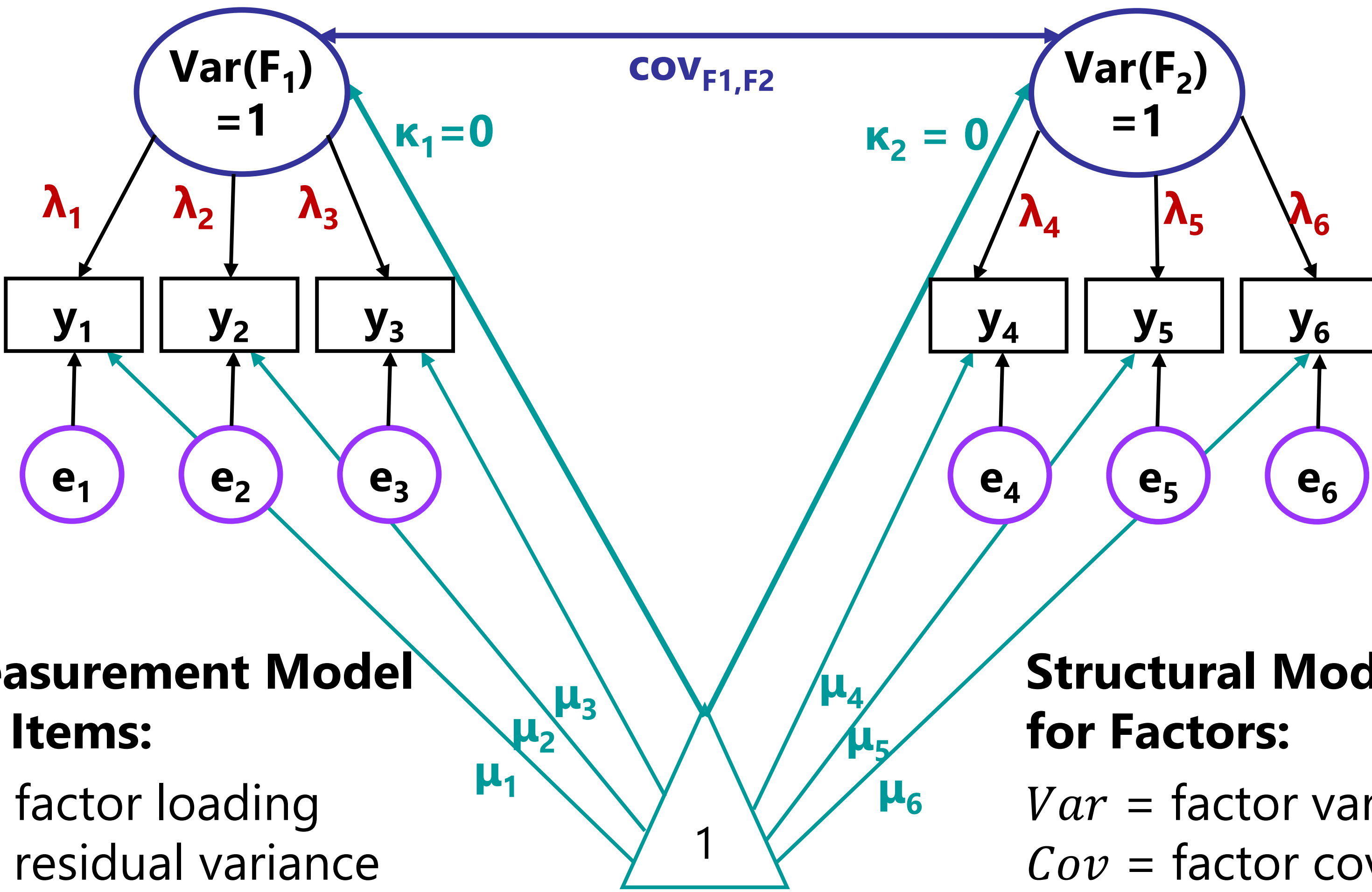
Residual covariances for same indicators across different occasions are usually included by default when possible

Separate factors are estimated for each occasion; covariances are always estimated to capture dependency of traits over time

FYI: A structural model in which all factor means, variances, and covariances are estimated is analogous to a "saturated means, unstructured variance model" for observed variables in MLM terms

Remember the CFA model? Let's start MI testing here….

We will begin with my preferred default of a fixed factor mean=0 and factor variance=1.



Measurement Model for Items:

λ = factor loading
e = residual variance
μ = intercept

Structural Model for Factors:

Var = factor variance
Cov = factor covariance
K = factor mean

Steps of Testing Invariance across Groups

- **Step 0 (optional): Omnibus test of equality of the overall indicator (item) covariance matrix across groups**
  - Do the covariances matrices differ between groups on the whole?
  - If not, game over. You are done. You have invariance. Congratulations.
  - Many people disagree about the necessity or usefulness of this omnibus test to begin testing invariance… *why might that be*?
  - People also differ in whether invariance should go from top-down or bottom-up directions… I favor bottom-up for the same reason.

- Let's proceed with an example with 2 factors, 6 indicators (3 per factor; simple structure), and 2 groups…
  - Total possible # parameters = $\frac{v(v+1)}{2} + v = \frac{6(6+1)}{2} + 6 = 27$ per group
  - So our COMBINED total possible $DF = 54$ across 2 groups

Step 1: Test "Configural" Invariance

- **Do the groups have the same general factor structure?**
- Same number of factors, same pattern of free/0 loadings
  → same conceptual definition of latent constructs
- In practice, begin by testing the factor structure within each group separately; hope they are "close enough"

- Then estimate group-specific models simultaneously, but **allow all estimated parameters to differ across groups**
  - This will be the baseline model for further comparisons
  - Absolute fit χ^2 and DF will be additive across groups (different group sample sizes will result in differential weighting of χ^2 across groups)

- This is as good fit as it gets! From here forward, our goal is to make model fit NOT WORSE by constraining parameters equal
  - That means if the configural model fits badly, game over…

Testing Invariance Constraints

- As before, we will test whether subtracting parameters worsens model fit via likelihood ratio (aka, −2ΔLL, χ^2) tests
  - Implemented via a direct difference in H_0 model χ^2 values most often, but this is only appropriate when using regular ML estimation
- MLR requires a modified version of this −2ΔLL test (see Mplus website): http://www.statmodel.com/chidiff.shtml
  - Is called a "rescaled likelihood ratio test" when you explain it
  - Includes extra steps to incorporate scaling factors (1.00 = regular ML)
  - I built you a spreadsheet for this…you're still welcome ☺
- If **removing parameters** (e.g., in invariance testing), H_0 model fit can get **worse OR not worse** (as indicated by smaller LL OR by larger −2LL and χ^2)
  - This is what we are doing in testing invariance!
- If **adding parameters** (e.g., in adding factors), H_0 model fit can get **better OR not better** (as indicated by larger LL OR by smaller −2LL and χ^2)

Testing Nested Models via $-2\Delta LL$

- Comparing nested models via a "**likelihood ratio test**" → $-$**2ΔLL** (MLR rescaled version)
  - 1. Calculate $-$**2ΔLL** = $-2 * (LL_{fewer} - LL_{more})$
  - 2. Calculate **difference scaling correction** =
    $$\frac{(\#parms_{fewer} * scale_{fewer}) - (\#parms_{more} * scale_{more})}{(\#parms_{fewer} - \#parms_{more})}$$
  - 3. Calculate **rescaled difference** = $-2\Delta LL$ / scaling correction
  - 4. Calculate **Δdf** = $\#parms_{more} - \#parms_{fewer}$
  - 5. **Compare rescaled difference to χ^2 with df = Δdf**
    - Add 1 parameter? $LL_{diff} > 3.84$, add 2 parameters: $LL_{diff} > 5.99$…
    - Absolute values of LL are meaningless (is relative fit only)
    - **Process generalizes to any other model estimated using maximum likelihood (such as IRT/IFA) that provides LL height for the H_0 model**

Note: Your LL will always be listed as the $\boldsymbol{H_0}$ (H_1 is for the saturated, perfectly fitting model)

Fewer = simpler model
More = more parameters

1. Configural Invariance Model:
Same Factor Structure; All Parameters Separate

Total DF across groups = 54 – 38 = 16 =
$$54-\left(12\mu+12\sigma_e^2+12\lambda+0\sigma_F^2+2\sigma_{F12}+0\kappa_F\right)=16$$

<u>Group 1 (subscript = item, group):</u>

- $y_{11s} = \mu_{11} + \lambda_{11}F_{1s} + e_{11s}$
- $y_{21s} = \mu_{21} + \lambda_{21}F_{1s} + e_{21s}$
- $y_{31s} = \mu_{31} + \lambda_{31}F_{1s} + e_{31s}$
- $y_{41s} = \mu_{41} + \lambda_{41}F_{2s} + e_{41s}$
- $y_{51s} = \mu_{51} + \lambda_{51}F_{2s} + e_{51s}$
- $y_{61s} = \mu_{61} + \lambda_{61}F_{2s} + e_{61s}$
- Estimated factor covariance, but both factor means = 0 and both factor variances = 1

<u>Group 2 (subscript = item, group):</u>

- $y_{12s} = \mu_{12} + \lambda_{12}F_{1s} + e_{12s}$
- $y_{22s} = \mu_{22} + \lambda_{22}F_{1s} + e_{22s}$
- $y_{32s} = \mu_{32} + \lambda_{32}F_{1s} + e_{32s}$
- $y_{42s} = \mu_{42} + \lambda_{42}F_{2s} + e_{42s}$
- $y_{52s} = \mu_{52} + \lambda_{52}F_{2s} + e_{52s}$
- $y_{62s} = \mu_{62} + \lambda_{62}F_{2s} + e_{62s}$
- Estimated factor covariance, but both factor means = 0 and both factor variances = 1

Step 2: Test "**Metric**" Invariance

- Also called "**weak** factorial invariance"
- Do the groups have the same **factor loadings**?
  - Each indicator is still allowed to have a different loading within each group → this is **not** a tau-equivalent model
  - Loadings for *same indicator* are constrained equal *across groups*

- Estimate all newly constrained factor loadings, but **fix all factor variances to 1 in the reference group** (freely estimate all factor variances in other group)
  - Why? Loadings for marker items (fixed=1 for identification) would be assumed invariant, and thus they could not be tested for invariance
  - This alternative specification allows us to evaluate ALL loadings and still identify the model (see Yoon & Millsap, 2007), which is BETTER

2. **Metric** Invariance Model:
Same Factor Loadings Only (saves 4 DF)

Total DF across groups = 54 – 34 = 20 $=$
$54 - \left(12\mu + 12\sigma_e^2 + \boldsymbol{6\lambda} + \boldsymbol{2\sigma_F^2} + 2\sigma_{F12} + 0\kappa_F\right) = 20$

Group 1 (subscript = item, group):

- $y_{11s} = \mu_{11} + \boldsymbol{\lambda_1 F_{1s}} + e_{11s}$
- $y_{21s} = \mu_{21} + \boldsymbol{\lambda_2 F_{1s}} + e_{21s}$
- $y_{31s} = \mu_{31} + \boldsymbol{\lambda_3 F_{1s}} + e_{31s}$
- $y_{41s} = \mu_{41} + \boldsymbol{\lambda_4 F_{2s}} + e_{41s}$
- $y_{51s} = \mu_{51} + \boldsymbol{\lambda_5 F_{2s}} + e_{51s}$
- $y_{61s} = \mu_{61} + \boldsymbol{\lambda_6 F_{2s}} + e_{61s}$
- **Both factor variances = 1 for identification**, factor covariance is estimated, but both factor means are STILL = 0

Group 2 (subscript = item, group):

- $y_{12s} = \mu_{12} + \boldsymbol{\lambda_1 F_{1s}} + e_{12s}$
- $y_{22s} = \mu_{22} + \boldsymbol{\lambda_2 F_{1s}} + e_{22s}$
- $y_{32s} = \mu_{32} + \boldsymbol{\lambda_3 F_{1s}} + e_{32s}$
- $y_{42s} = \mu_{42} + \boldsymbol{\lambda_4 F_{2s}} + e_{42s}$
- $y_{52s} = \mu_{52} + \boldsymbol{\lambda_5 F_{2s}} + e_{52s}$
- $y_{62s} = \mu_{62} + \boldsymbol{\lambda_6 F_{2s}} + e_{62s}$
- **Both factor variances are now estimated,** factor covariance is still estimated, but both factor means are STILL = 0

2. **Metric** Invariance Model

- Compare metric invariance to configural invariance model: **Is the model fit not worse** ($-2\Delta LL$ not significant)?
  - Check that factor variances are fixed to 1 in the reference group only: they should be freely estimated in the other group, otherwise you are imposing a structural constraint (that groups have same variability) too
  - Otherwise, inspect the **modification indices** to see if there are any indicators whose loadings want to differ across groups instead
  - Re-estimate the model after releasing one loading at a time, starting with the largest modification index, and continue until your partial metric invariance model is **not worse** than the configural model

- Do you have partial metric invariance (1+ loading per factor)?
  - Your trait is (sort of) measured in the same way across groups
  - If not, it doesn’t make sense to evaluate how relationships involving the factor differ across groups (because the factor itself differs)
  - Even if full invariance holds, check the modification indices anyway (because individual discrepancies can be washed out in overall test)

Step 3: Test "**Scalar**" Invariance

- Also called "**strong** factorial invariance"
- Do the groups have the same **indicator intercepts**?
  - Each indicator is still allowed to have a different intercept, but intercepts for same indicator are constrained equal across groups
  - Indicators that failed metric invariance traditionally do not get tested for scalar invariance (single-elimination), but they could be
  - Scalar invariance is required for factor mean comparisons!

- Previous (partial) metric invariance model is starting point
- Estimate all newly constrained intercepts, but
  fix the factor means to 0 in the reference group
  (**free** the factor means in the other group)
  - Why? Intercepts for marker items (if fixed=0 for identification) would be assumed invariant, and thus they could not be tested

3. Scalar Invariance Model:
Same Factor Loadings + Same Intercepts (saves +4 DF)

Total DF across groups = 54 – 30 = 24 $=$
$54 - \left(\boldsymbol{6\mu} + 12\sigma_e^2 + 6\lambda + 2\sigma_F^2 + 2\sigma_{F12} + \boldsymbol{2\kappa_F}\right) = 24$

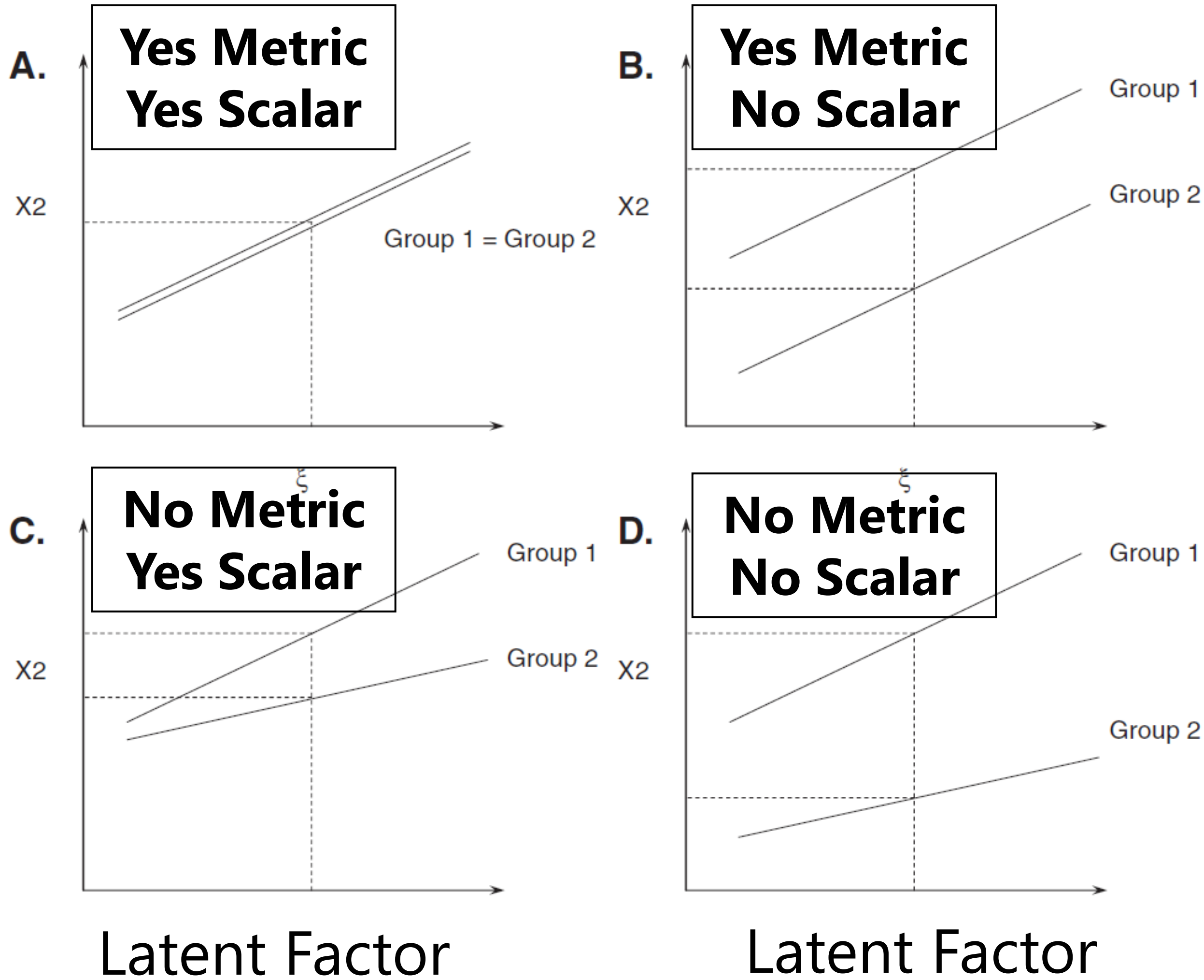
<u>Group 1 (subscript = item, group):</u>

- $y_{11s} = \boldsymbol{\mu_1} + \lambda_1 F_{1s} + e_{11s}$
- $y_{21s} = \boldsymbol{\mu_2} + \lambda_2 F_{1s} + e_{21s}$
- $y_{31s} = \boldsymbol{\mu_3} + \lambda_3 F_{1s} + e_{31s}$
- $y_{41s} = \boldsymbol{\mu_4} + \lambda_4 F_{2s} + e_{41s}$
- $y_{51s} = \boldsymbol{\mu_5} + \lambda_5 F_{2s} + e_{51s}$
- $y_{61s} = \boldsymbol{\mu_6} + \lambda_6 F_{2s} + e_{61s}$
- **Both factor variances fixed to 1, both factor means fixed to 0 for identification**, factor covariance is still estimated

<u>Group 2 (subscript = item, group):</u>

- $y_{12s} = \boldsymbol{\mu_1} + \lambda_1 F_{1s} + e_{12s}$
- $y_{22s} = \boldsymbol{\mu_2} + \lambda_2 F_{1s} + e_{22s}$
- $y_{32s} = \boldsymbol{\mu_3} + \lambda_3 F_{1s} + e_{32s}$
- $y_{42s} = \boldsymbol{\mu_4} + \lambda_4 F_{2s} + e_{42s}$
- $y_{52s} = \boldsymbol{\mu_5} + \lambda_5 F_{2s} + e_{52s}$
- $y_{62s} = \boldsymbol{\mu_6} + \lambda_6 F_{2s} + e_{62s}$
- **Both factor variances estimated, both factor means estimated to become mean differences,** and factor covariance is still estimated

Implications of Non-Invariance



Without metric invariance: Because unequal loadings implies non-parallel slopes, the intercept can differ as a result. The size of the difference depends on where trait=0.

This is why scalar invariance is often not tested if metric invariance fails for a given indicator.

Figure 7.4 borrowed from Brown (2015)

3. **Scalar** Invariance Model

- Compare scalar invariance to last metric invariance model: **Is the model fit not worse** ($-2\Delta LL$ not significant)?
  - Check that factor means are fixed to 0 in reference group only: they should be freely estimated in the other group, otherwise you are imposing a structural constraint (groups have same means) too
  - Otherwise, inspect the **modification indices** to see if there are any indicators whose intercepts want to differ across groups instead
  - Re-estimate the model after releasing one intercept at a time, starting with the largest modification index, and continue until your partial scalar invariance model is **not worse** than last metric model

- Do you have partial scalar invariance (1+ intercept per factor)?
  - Your trait is (sort of) responsible for mean differences across groups
  - If not, it doesn't make sense to evaluate factor means differs across groups (because something else is causing those differences)
  - Even if full invariance holds, check the modification indices anyway

Step 4: Test **Residual Variance** Invariance

- Also called "**strict** factorial invariance"
- Do the groups have the same **residual variances**?
  - Each indicator is still allowed to have a different residual variance (i.e., this is not a parallel items model), but residual variances for *same indicator* are constrained equal *across groups*
  - Indicators that failed scalar invariance traditionally do not get tested for residual variance invariance (although they could be)
  - Residual invariance is of debatable importance, because it means that whatever created "not the factor" does not differ by group
  - Equal residual variances are commonly misinterpreted to mean "equal reliabilities"—this is ONLY the case if the factor variances are the same across groups, too (stay tuned)

- Without residual covariances, this is the last step of "measurement invariance" (otherwise they are last)

4. Residual Invariance Model:

+ Same Residual Variances (saves +6 DF)

Total DF across groups = 54 – 24 = 30 $=$
$$54 - \left(6\mu + \boldsymbol{6\sigma_e^2} + 6\lambda + 2\sigma_F^2 + 2\sigma_{F12} + 2\kappa_F\right) = 30$$

Group 1 (subscript = item, group):

- $y_{11s} = \mu_1 + \lambda_1 F_{1s} + \boldsymbol{e_{1s}}$
- $y_{21s} = \mu_2 + \lambda_2 F_{1s} + \boldsymbol{e_{2s}}$
- $y_{31s} = \mu_3 + \lambda_3 F_{1s} + \boldsymbol{e_{3s}}$
- $y_{41s} = \mu_4 + \lambda_4 F_{2s} + \boldsymbol{e_{4s}}$
- $y_{51s} = \mu_5 + \lambda_5 F_{2s} + \boldsymbol{e_{5s}}$
- $y_{61s} = \mu_6 + \lambda_6 F_{2s} + \boldsymbol{e_{6s}}$
- **Both factor variances fixed to 1, both factor means fixed to 0 for identification**, factor covariance is still estimated

Group 2 (subscript = item, group):

- $y_{12s} = \mu_1 + \lambda_1 F_{1s} + \boldsymbol{e_{1s}}$
- $y_{22s} = \mu_2 + \lambda_2 F_{1s} + \boldsymbol{e_{2s}}$
- $y_{32s} = \mu_3 + \lambda_3 F_{1s} + \boldsymbol{e_{3s}}$
- $y_{42s} = \mu_4 + \lambda_4 F_{2s} + \boldsymbol{e_{4s}}$
- $y_{52s} = \mu_5 + \lambda_5 F_{2s} + \boldsymbol{e_{5s}}$
- $y_{62s} = \mu_6 + \lambda_6 F_{2s} + \boldsymbol{e_{6s}}$
- **Both factor variances estimated, both factor means estimated to become mean differences,** and factor covariance is still estimated

4. **Residual Variance** Invariance Model

- Compare residual invariance to last scalar invariance model: **Is the model fit not worse** (-2ΔLL not significant)?
  - Otherwise, inspect the **modification indices** to see if there are any indicators whose residual variances want to differ across groups
  - Re-estimate the model after releasing one residual variance at a time, starting with the largest modification index, and continue until your partial residual invariance model is **not worse** than last scalar model
- Do you have partial residual variance invariance (1+ residual variance per factor)?
  - Your groups have the same amount of "not the factor" variance in each item (so that's good, I guess???)
  - Even if full invariance holds, check the modification indices anyway
  - Also assess **any residual covariances** across groups if you have those
- Your (partial) residual invariance model is the new baseline for assessing structural invariance…

Testing **Structural** Invariance

- Are the **factor variances** the same across groups? (+1 DF per factor)
  - Fix each factor variance in the alternative group to 1 (as in the ref group)
  - Is model fit worse? If so, the groups differ in their factor variances
- Are the **factor covariances** the same across groups? (+1 DF per pair)
  - Fix each factor covariance equal across groups, is model fit worse?
  - Factor correlation will only be the same across groups if the factor variances are the same, too (if factor variances differ, then factor covariance will, too)
- Are the **factor means** the same across groups? (+1 DF per factor)
  - Fix each factor mean in the alternative group to 0 (as in the ref group)
  - Is model fit worse? If so, the groups differ in their factor means
  - Btw, the Wald test for the factor means already tell you this (difference from 0)
- **It is not a problem if structural invariance doesn't hold!**
  - Given measurement invariance, this is a substantive issue about differences in the amounts and relations of the latent traits (and that's ok)
  - Might stop at measurement invariance for testing RQs involving the traits

Summary: Invariance Testing in CFA

- In CFA: Testing invariance has two distinct parts:
  - **Measurement invariance**: Is your construct being measured in the same way by the indicators across groups/time?
    - Hope for at least "partial" invariance… otherwise, it's game over for comparing how factors relate (or mean differences) across groups/time
  - **Structural invariance**: Do your groups/times differ in their distribution and/or means of the construct? Let's find out!
    - Structural differences are real and interpretable differences given measurement invariance of the constructs

- In IFA: Still called "testing invariance"
  - Conducted similarly (but not exactly the same) in Mplus
- In IRT: Now called testing "differential item functioning"
  - With different names and rules, not directly tested in Mplus

Differential Item Functioning (DIF)

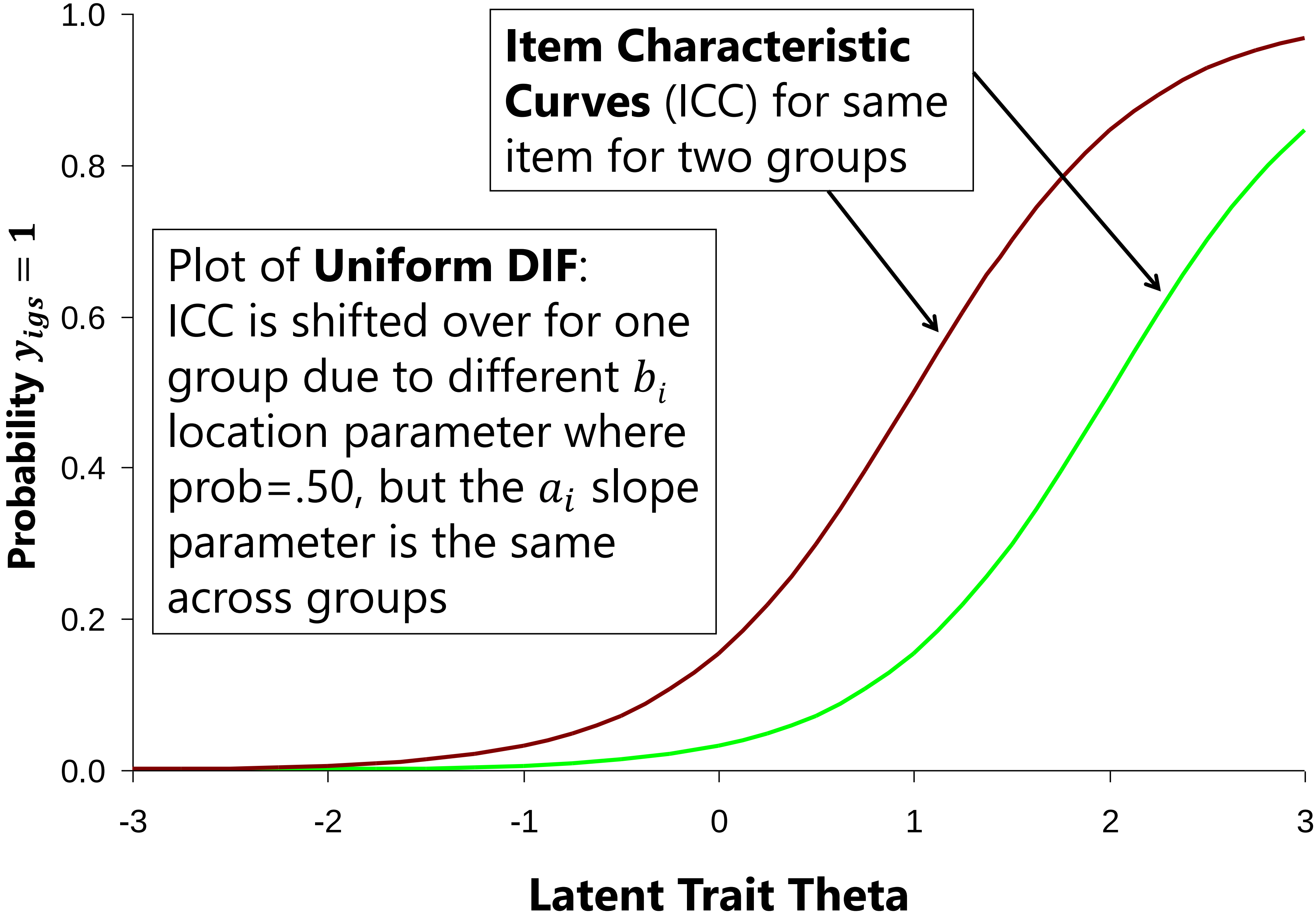
- In IRT (model with a_i discrimination and b_i difficulty), **measurement NON-invariance = DIF**
  - Note the inversion: Measurement Invariance = *Non*-DIF
    Measurement *Non*-Invariance = DIF
- An item has "**DIF**" when persons with equal trait amounts, but from different groups, have different expected item responses
- An item has "**non-DIF**" if persons with the same trait amount have the same expected item response, regardless of group
- DIF can be examined across groups, over time, over repeated conditions, etc., the same as in CFA/IFA
  - Independent groups? Multiple-group model
  - Dependent "groups"? One factor per "group" in same model

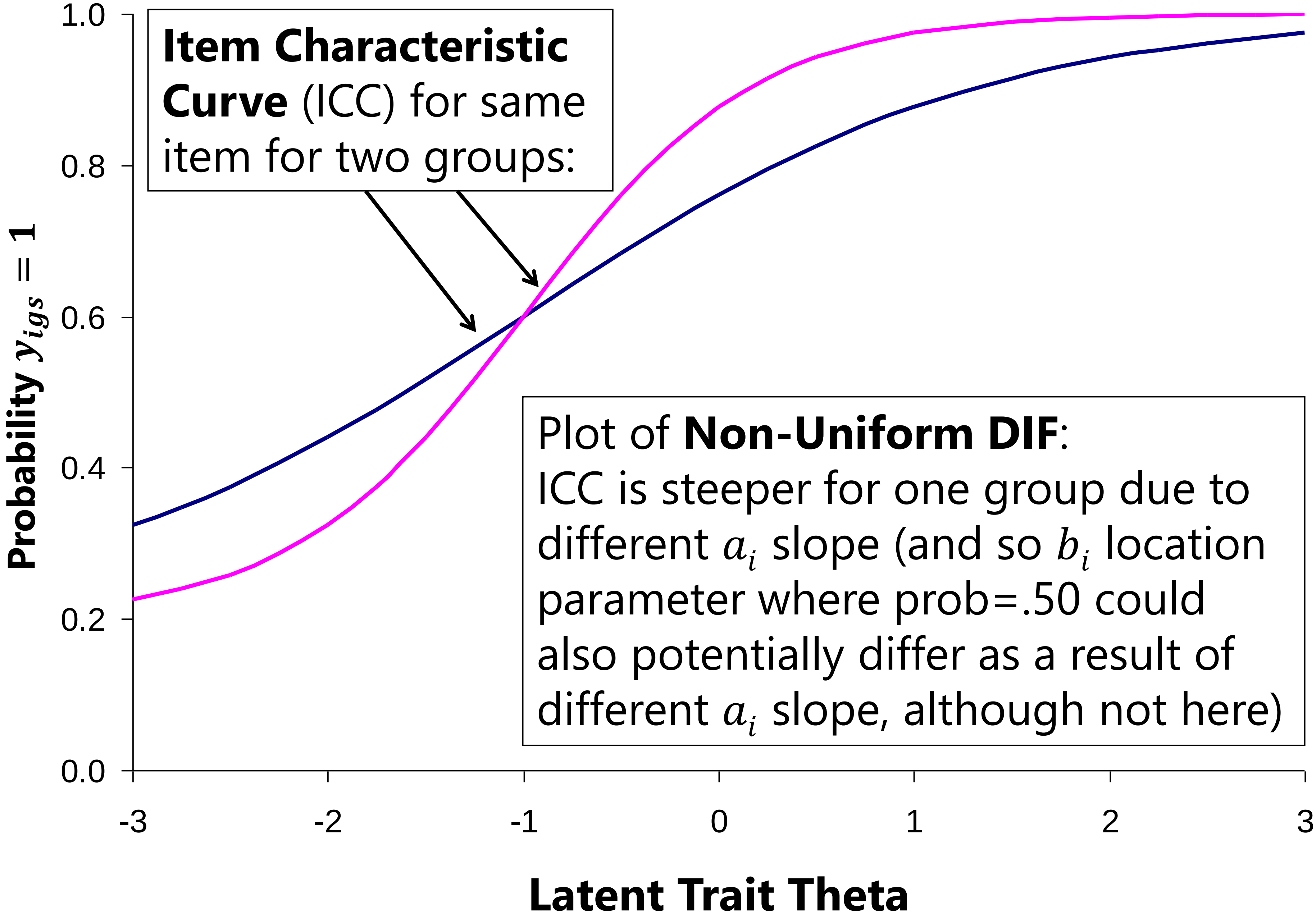
2 Types of DIF (as described in IRT)

- "**Uniform DIF**" → Analogous to scalar NON-invariance
  - **IRT b_i parameters differ across groups**
  - Item is systematically more difficult/severe for members of one group, *even for persons with the same amount of the theta trait*
  - Example: "How often do you cry?" → Would men and women *with the same amount of depression* have the same expected item response?

- "**Non-Uniform DIF**" → Analogous to metric NON-invariance
  - **IRT a_i (and possibly b_i) parameters differ across groups**
  - Item is systematically more related to theta for members of one group → higher discrimination (item "works better" in one group than another)
  - Group-related shift in item difficulty is not consistent across the trait

- What about **residual variance invariance**? It depends:
  - **Doesn't exist in ML**: no estimated error variance (is logit=3.29 or probit=1.00 for y^*)
  - Will exist **in WLSMV after constraining** loadings and thresholds, but not before…


Item Characteristic Curves (ICC) for same item for two groups
Plot of Uniform DIF:
ICC is shifted over for one group due to different b_i location parameter where prob=.50, but the a_i slope parameter is the same across groups
Probability $y_{igs} = 1$
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
Latent Trait Theta


Item Characteristic Curve (ICC) for same item for two groups:
Plot of **Non-Uniform DIF**:
ICC is steeper for one group due to different a_i slope (and so b_i location parameter where prob=.50 could also potentially differ as a result of different a_i slope, although not here)
Probability $y_{igs} = 1$
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
Latent Trait Theta

Testing Measurement Invariance for Categorical Outcomes

- 2 versions of model for polytomous outcomes in Mplus:
  - IRT: $\text{Logit or Probit}(y_{kis} = 1) = a_i(\theta_s - b_{ki})$
  - IFA: $\text{Logit or Probit}(y_{kis} = 1) = -\tau_{ki} + \lambda_i\theta_s$
    - Logit or Probit in ML; only Probit in WLSMV

The k thresholds divide the C item responses into $C-1$ cumulative binary submodels ($y = 0$ if lower, $y = 1$ if higher)

- Mplus estimates the IFA τ_{ki} and λ_i parameters, then *converts* to the IRT a_i and b_{ki} parameters for binary outcomes **after rescaling trait** (M=0, SD=1)
  - Tests of measurement invariance are thus specifically for τ_{ki} and λ_i, not a_i and b_{ki}
  - So Mplus and Lavaan do not directly test "DIF" for a_i and b_i parameters
- **IFA τ_{ki} and λ_i parameters are held directly invariant, not IRT a_i and $\mathrm{b_i}$**
  - So even if λ_i factor loadings are invariant across groups, the IRT a_i discriminations given by Mplus will still differ across groups due to differences in their theta variances (but you can calculate the invariant a_i parameters yourself via MODEL CONSTRAINT)
  - Likewise, even if τ_{ki} thresholds are invariant across groups/time, Mplus IRT b_i difficulty parameters will still differ due to their rescaling of the trait (but you can fix this)

Review: From IFA to IRT

IFA with "easiness" **intercept** $\boldsymbol{\mu_i}$: **Logit or Probit** $\boldsymbol{y_{is} = \mu_i + \lambda_i F_s \quad \mu_i = -\tau_i}$

IFA with "difficulty" **threshold** $\boldsymbol{\tau_i}$: **Logit or Probit** $\boldsymbol{y_{is} = -\tau_i + \lambda_i F_s}$

IFA model with "difficulty" thresholds can be written as a **2-PL IRT Model:**

IRT model: **IFA model:**

$$\textbf{Logit or Probit } \boldsymbol{y_{is} = a_i(\theta_s - b_i) \;=\; -\underbrace{a_i b_i}_{\tau_i} + \underbrace{a_i}_{\lambda_i}\theta_s}$$

$\boldsymbol{a_i}$ = discrimination
$\boldsymbol{b_i}$ = difficulty
$\boldsymbol{\theta_s}$ = $\boldsymbol{F_s}$ latent trait

Convert IFA to IRT:

$$a_i = \lambda_i * \sqrt{\text{Theta Variance}}$$

$$b_i = \frac{\tau_i - (\lambda_i * \text{Theta Mean})}{\lambda_i * \sqrt{\text{Theta Variance}}}$$

Convert IRT to IFA:

$$\lambda_i = \frac{a_i}{\sqrt{\text{Theta Variance}}}$$

$$\tau_i = a_i b_i + \frac{a_i * \text{Theta Mean}}{\sqrt{\text{Theta Variance}}}$$

Note: These formulas rescale a_i and b_i so that theta M=0, VAR=1

If you don't want to rescale theta, use M=0 and VAR=1 to keep your current scale

Invariance Testing in Mplus

- **<u>IFA models using Full-Information MML:</u>** Mplus must be tricked into estimating a multiple-group models (e.g., here, by group):
  - VARIABLE: CLASSES=group(2); KNOWNCLASS = group (female=0 1);
  - ANALYSIS: TYPE = MIXTURE; ESTIMATOR = ML; ALGORITHM = INTEGRATION;
  - MODEL: %OVERALL% (… model for reference group listed here)

    %group#2% (… model for alternative group goes here)

- **<u>IFA models using Limited-Information WLSMV:</u>** Mplus directly estimates multiple-group models, with a few useful other benefits
  - Faster estimation if you have multiple factors/thetas (but assumes MCAR!)
  - DIFFTEST does nested model comparisons for you (still going for "not worse")
  - Can get modification indices to determine sources of non-invariance
  - Can test differences in residual variances (in THETA parameterization)

- In WLSMV, the same category responses must be observed for each group, otherwise you cannot test the item thresholds
  - But using MML, you can estimate more thresholds in one group than another

Configural Invariance Baseline Model for Categorical Outcomes (2 Groups)

- **Factor variances**: fixed to 1 in both groups
- **Factor covariances**: if any, free in both groups
- **Factor means**: fixed to 0 in both groups

We use the same configural model identification as in CFA for simplicity (but it doesn't really matter how it's identified here)

- **Factor loadings**: all freely estimated (so each can be tested later)
  - Remember: the IRT a_i parameters Mplus gives you will still vary across groups even after loadings are constrained because of group differences in theta variance
- **Item Thresholds**: all freely estimated (so each can be tested later)
  - Remember: the IRT b_{ki} parameters Mplus gives you will still vary across groups even after thresholds are constrained because of group differences in theta mean and theta variance
- **Fix all residual variances=1 in all groups <u>if using WLSMV</u>**
  - Groups will eventually be allowed to differ in both factor variance and "error variance" (proxy for total variation in WLSMV models; is a "slop" parameter)

Sequential Invariance Models
Note: In WLSMV, save for DIFFTEST at each step!

- **Step 1**: Fit same baseline configural invariance model across groups
  - Should be "close enough" factor structures, otherwise game over
  - Alternative group is allowed different loadings and thresholds
- **Step 2 (Metric-ish)**: Constrain all loadings equal, but free the factor variances in alternative group—is fit worse than the configural model?
  - If using WLSMV, check MODINDICES to see misfit of constraints; release problematic constrained loadings <u>one at a time</u>; check fit against configural model to see if it's not worse yet
- **Step 3 (Scalar-ish)**: Constrain thresholds equal for items (that passed metric) but free factor means in alternative group—is fit worse than the metric model?
  - If using WLSMV, check MODINDICES to see misfit of constraints; release problematic constrained thresholds <u>one item at a time</u>; check fit against metric model to see if it's not worse yet
  - If using WLSMV, MODINDICES may want the "intercept" free, but this is not possible to do in IRT/IFA, so focus on problematic (non-invariant) item thresholds instead
  - Reasonable people disagree: Mplus recommends doing steps 2 and 3 in one step because loadings and thresholds are dependent; me and others disagree (e.g., "uniform DIF" in IRT)

Sequential Invariance Models

Note: In WLSMV, save for DIFFTEST at each step!

- **Step 4**: *WLSMV only*: Test if residual variances for items (that passed scalar) in alternative group ≠1 → differ from reference group (of residual variance = 1)
  - Differences in residual variances between groups are not identified until you have at least some of the loadings and thresholds constrained across groups
  - Consequently, **this test proceeds backwards**: first estimated is the "bigger" non-invariant residual variance model, second estimated is the "smaller" original scalar invariance model with new name (in which residual variances were fixed to 1 for all items for all groups)
  - Differential residual variances can be a proxy for group differences in overall variability (or slop), but this model may not always converge (if it doesn't, just skip this step, but note doing so)

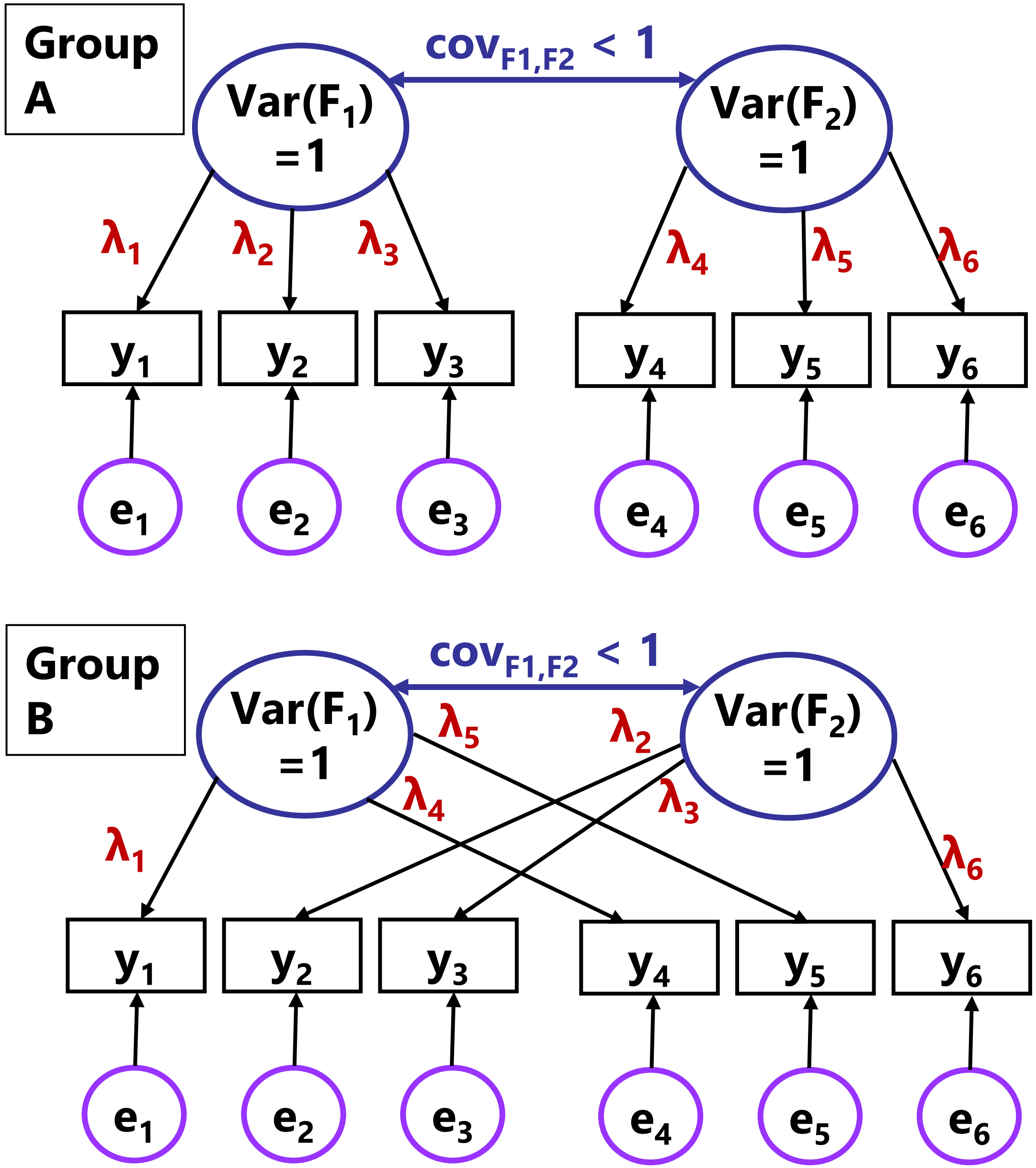
- **Steps 5, 6, 7:** Test Structural Invariance (just like before in CFA):
  - Constrain equal across groups in sequential models: factor variances, then factor covariances, and then factor means (hold equal to 0) to test for "real" structural group differences
  - Same story as in CFA: Only if you have at least partial measurement invariance can structural group/time/condition differences be meaningfully interpreted

- Factors/Thetas all should have a multivariate normal distribution no matter what measurement model was used to create them… so now we can do SEM!

What If I Don't Have Groups?

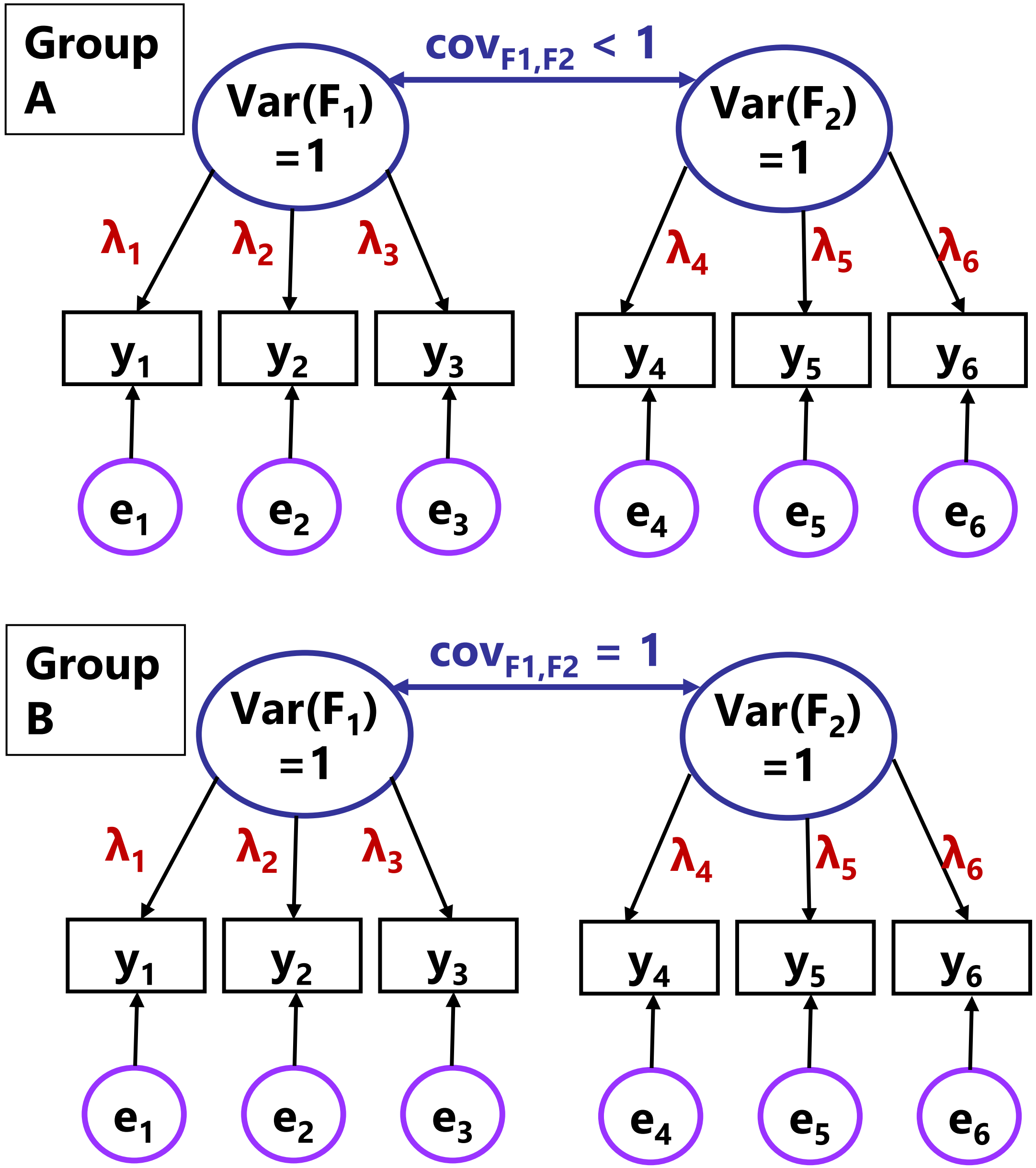
- Invariance/DIF testing traditionally occurs across levels of categorical (grouping) variables, but it doesn't have to
- If you want to examine invariance across ranges of quantitative predictors instead, you can do so via "**moderated nonlinear factor analysis**"
  - See Bauer & Hussong (2009) and Curran et al. (2014) for examples
- Key idea: Allow loadings and intercepts/thresholds to differ systematically across persons using a linear model
  - e.g., Per item: $\lambda = \beta_0 + \beta_{age}(age_s) + \beta_{group}(group_s)$
  - Can be done in Mplus using MODEL CONSTRAINT (need to add variables used in linear model to CONSTRAINT= in VARIABLE section)
  - Of course there is an R package for it, too

Configural Invariance: Game Over



- **Configural invariance** requires the same factor structure across groups with respect to how the items relate to their factors
- **The factors need to plausibly measure the same latent constructs, otherwise it's game over**
- Example: Groups A and B both have two factors, but they are measured using different items
- So it doesn't make sense to test if the items relate the same way across groups to different factors!

Configural Invariance: Work-Around



- **Configural invariance** requires the same factor structure across groups with respect to how the items relate to their factors
- **Different numbers of factors could be ok IF you can find a way to phrase the models equivalently**

- Example: Group A needs 2 factors, but Group B only needs 1 factor
- Solution: Fit 2 factors in both groups, but **let the factor covariance differ** (may need to constrain it to prevent NPD solutions)