

Item Response Theory Part 4: Differential Item Functioning

- Today's topics:
 - Reviewing measurement invariance
 - Types of Differential Item Functioning (DIF)
 - Baseline configural model specification
 - Testing for DIF
 - Wrapping Up...

Measurement Invariance

- Remember “measurement invariance”?
 - Measurement invariance holds when two persons in different groups have the same expected raw item responses given the same level of the latent trait
 - Otherwise, if group differences are obtained, is it because...
 - Indicators relate differently to the latent trait (loadings, intercepts) across groups? *Lack of Measurement Invariance*
 - Groups really are different in their levels, distributions, or relations among latent traits? *Lack of Structural Invariance*

Steps in Assessing Invariance

- Measurement Invariance:
 - Overall covariance matrix (maybe)
 - Test Loadings → “metric” or “weak” invariance
 - Test Intercepts → “scalar” or “strong” invariance
 - Test Error variances → “strict” invariance
- Assuming at least partial measurement invariance holds, then we can assess Structural Invariance:
 - Factor variances, then covariances if variances are equal
 - Distributions of latent trait and relations among them
 - Factor means
 - Level of the latent trait

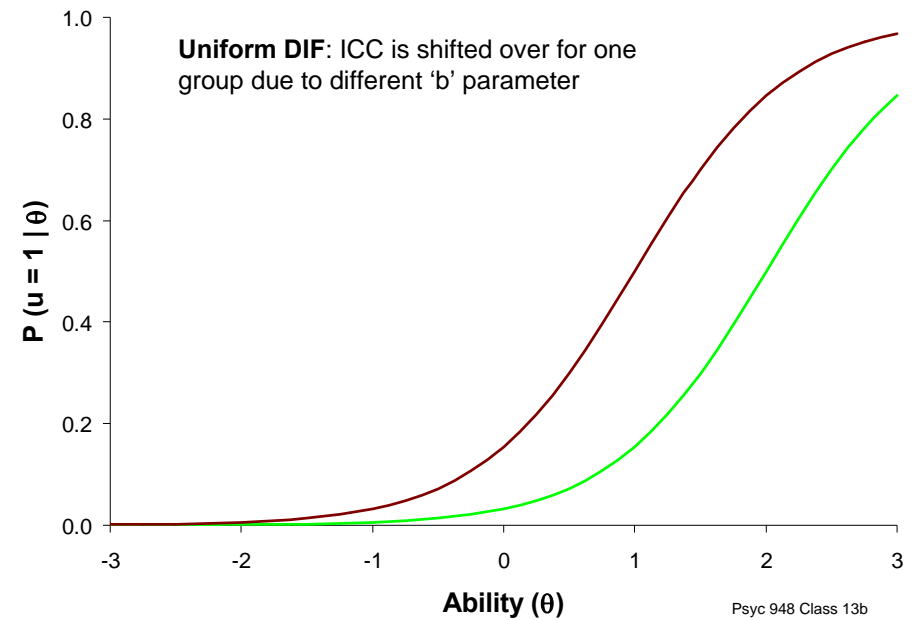
Measurement Invariance in IRT

- In IRT, Measurement Non-Invariance is called **“Differential Item Functioning” (DIF)**
 - We expect people to vary in their levels of the latent trait.
 - It's when people with the same latent trait levels have different expected item responses that is the potential problem.
 - Note the odd language: Measurement Invariance = Non-DIF
Measurement Non-Invariance = DIF
 - An item is labeled with “DIF” when persons with equal ability, but from different groups, have an unequal probability of item success.
 - An item is labeled as “non-DIF” if persons having the same ability have equal probability of getting the item correct, regardless of group membership.
 - Group membership (e.g., gender, ethnicity) should not *differentially* impact success across items.

2 Flavors of DIF

- “Uniform DIF”
 - Analogous to scalar (intercept) invariance
 - CFA “Thresholds” (or IRT “b” parameters) differ across groups
 - Item is systematically more difficult for members of one group, *even after controlling for theta*
 - Example: “I cry a lot”
 - Would men and women with equal latent levels of depression both have the same expected item response??

Psyc 948 Class 13b
5 of 12

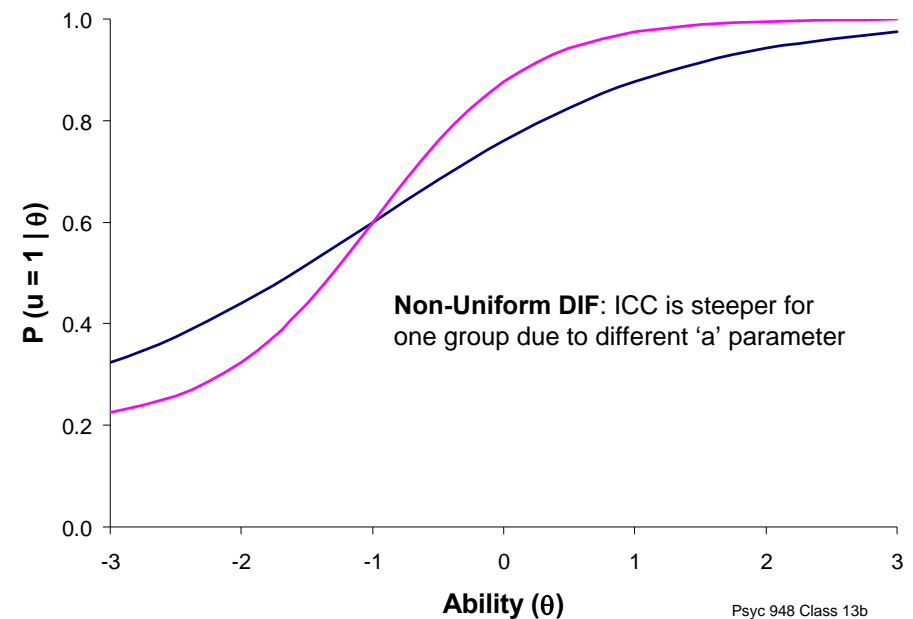


Psyc 948 Class 13b
6 of 12

2 Flavors of DIF

- “Non-Uniform DIF”
 - Analogous to metric (loading) invariance
 - CFA “Loadings” (or IRT “a” parameters) (*and possibly “threshold” or “b” parameters too*) differ across groups
 - Item is systematically more related to theta (“works better”) for members of one group
 - Therefore, shift in item difficulty is not consistent across the ability continuum

Psyc 948 Class 13b
7 of 12



Psyc 948 Class 13b
8 of 12

How to Test for DIF using Mplus

- Mplus does not provide multiple group models using ML with numeric integration, so we use **WLSMV** instead
 - Faster estimation, requires fewer people than ML
 - Can get direct nested comparisons using DIFFTEST
- Model parameters to be tested for DIF (non-invariance within a item factor model framework) include:
 - λ_i Factor loadings (like “ a_i ” discriminations) → “non-uniform DIF”
 - τ_{ic} Thresholds (like “ b_{ic} ” difficulties) → “uniform DIF”
 - e^2 Residual variances → (no label; only in IFA context)
 - Structural parameters (factor mean, variances, covariances)
- IRT parameters are not directly held invariant, only IFA

Configural Baseline Model for Categorical Outcomes and 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
- **Factor loadings:** all free (so can be tested later)
 - Discriminations will still vary across groups even after loadings are constrained
- **Item Thresholds:** all free (so can be tested later)
 - Difficulties will still vary across groups even after thresholds are constrained
- **Fix all residual variances=1** in in both groups
 - Groups will eventually be allowed to differ in both factor variance and “error variance” (proxy for total variation in WLSMV models)

Sequential Models for Testing DIF

Note: Save results at each step!

- Step 1: Fit baseline configural model across groups
 - Should be ‘close enough’ factor models, otherwise game over
 - Alt group is allowed different loadings, thresholds, & residual variances
- Step 2: Constrain loadings equal across groups and free factor variances in alternative group – did fit get worse?
- Step 3: Constrain thresholds equal across groups and free factor mean in alternative group – did fit get worse?
- Step 4: New model: RELEASE constraints on residual variances in alternative group, save results, then compare against step 3
- Steps 5, 6, 7 test Structural Invariance:
 - Constrain equal across groups: factor variances, factor covariances, and then factor means (equal to 0) to test for “real” group differences

DIF: Wrapping Up...

- The analog to measurement invariance in CFA models is differential item functioning in IFA and IRT models...
 - Different difficulties (tested via thresholds)? “Uniform DIF”
 - Different discriminations (tested via loadings)? “Non-uniform DIF”
 - Can also test different residual variances in IFA framework
- Process of nested model comparisons to do so in WLSMV can be tricky in multiple group models...
 - First model must be the one with MORE parameters, save results
 - Second model has FEWER parameters, compare with first
 - Reversal of this process necessary to test freed residual variances (which are not separately identifiable before testing thresholds)