Measurement Equivalence/ Invariance

PSY 600K
Dr. Alyssa Gibbons
April 17, 2018

Agenda

- Measurement equivalence / invariance
  - Why we worry about it
  - What it is
  - How to test it

Measurement Equivalence/Invariance

- When we want to **compare groups** on a latent construct, we need to be sure we are measuring the latent construct the same way in both groups.
  - Otherwise, our comparison may not mean much.
- Cross-cultural research provides strong examples of cases where we might worry about this:
  - E.g., studying martial satisfaction among U.S. and Afghan women.
  - Why should you be cautious about comparing means here?

Challenges in Comparison

- When is this an issue?
  - **Translated** measures
  - Culture (even when language is not a barrier).
  - Gender (sometimes)
  - Translated media – e.g., online vs. paper-and-pencil
  - And… ?
- We can't *assume* groups are different… and we can't assume they aren't.
  - Need a way to separate true differences on the underlying construct from differences in the way the groups use or interpret the measurement instrument.
  - Two major strategies: MGCFA and IRT

MGCFA

- Multiple Group Confirmatory Factor Analysis
  - Goes beyond the individual item-level bias indices we talked about last week (and works better, too!).
  - Considers all items together in a factor analytic framework.
- Essentially, tests whether the same CFA model holds for two (or more) groups.
  - We do this by **constraining** parameters to be equal across groups – does the model still fit?

Formally

- Vandenberg & Lance (2000) use a very formal notation for CFA. The **model** is the same, just the notation is different!
  - This notation comes from LISREL (Jöreskog & Sörbom).
- We can write the model for item responses as:
  - $\mathbf{X}_k^g = \boldsymbol{\tau}_k^g + \boldsymbol{\Lambda}_k^g \boldsymbol{\xi}^g + \boldsymbol{\delta}_k^g$
  - $\mathbf{X}_k^g$ are the item responses (observed data)
  - $\boldsymbol{\tau}_k^g$ are the item means (also called *thresholds*). We don't always model these in normal CFA because they don't affect the covariance matrix – but we may care about them in measurement equivalence because they are our item difficulty parameters!
  - $\boldsymbol{\Lambda}_k^g$ are the factor loadings
  - $\boldsymbol{\xi}^g$ are the true scores (note that there isn't an item subscript here)
  - $\boldsymbol{\delta}_k^g$ are the uniquenesses (residual variances)

Formally, cont.

- So if our model is:
  - $\mathbf{X}_k^g = \tau_k^g + \boldsymbol{\Lambda}_k^g \boldsymbol{\xi}^g + \boldsymbol{\delta}_k^g$
- We can then write the covariance matrix as:
  - $\boldsymbol{\Sigma}^g = \boldsymbol{\Lambda}_X^g \boldsymbol{\Phi}^g \boldsymbol{\Lambda}_X^{g\prime} + \boldsymbol{\Theta}_{\delta k}^g$
- Why are we bothering?
  - It's good to be able to translate!
  - Also – all the *gs* in this model indicate that these are parameters that *might* be specific to a particular *group.*
  - **Every parameter matrix with a *g* is a set of parameters that could potentially vary from one group to another.**

ME/I Hypotheses

- In ME/I analysis, we can (and will!) test whether each piece of the model is equivalent across groups.
  - Each of these tests has a name.
- $\boldsymbol{\xi}^g$ is the same across groups – the construct has the same number of factors in all groups.
  - **Configural invariance**.
- $\boldsymbol{\Lambda}_k^g$ is the same across groups – the items load on the same factors and to the same extent across groups.
  - **Metric invariance.**

More ME/I Hypotheses

- $\boldsymbol{\tau}_k^g$ is the same across groups – the items have the same intercepts (difficulty) across groups.
  - **Scalar invariance.**
- $\boldsymbol{\Theta}_{\delta k}^g$ is the same across groups – the item uniquenesses or residual variances are the same in all groups.
  - **Invariance of uniquenesses.**
- $\boldsymbol{\Phi}^g$ is the same across groups – the factor variances and covariances are the same in all groups.
  - Actually, we can break this into two tests (variances & covariances.
- And we can also test whether the **factor means** are equal.

How to Do It

- **Free** parameters vs. **constrained** parameters.
  - Fit the model to both groups simultaneously, but allow the particular parameters you are interested in to be **different** in each group (free).
  - Then fit the model again, but require that those parameters be **equal** (constrained) across the groups.
    - Not specifying a value for those parameters, just saying they need to be equal.
- One-item example:
  - Test 1 (free):
    - x = .735F + .543 in Group 1, x = .541F + .489 in Group 2
  - Test 2 (constrained):
    - x = .601F + .543 in Group 1, x = .601F + .489 in Group 2

Model Comparisons

- We constrain one set of parameters at a time.
  - Adding constraints as we go – keep constraints that worked from previous steps.
- At each step, we test whether the fit of the constrained model is significantly worse than the fit of the free model.
  - Remember that the free model will always fit better… the question is how much better.
  - Is it reasonable to use 1 set of parameters to describe both groups?
- How do we test it?
  - Chi-square difference test
  - Change in CFI of .01 or greater (Cheung & Rensvold, 2001).
    - Implies a big enough difference to care about!
    - Based on simulation data – not totally arbitrary.

Order of Operations

- #1. First, test a **fully restricted model** – that covariance matrices are equal.
  - If so, you have ME/I! You're done.
  - If not, proceed to Step 2.
- #2. Test **configural invariance**.
  - Same # of factors, same items on each factor.
  - If you can't fit the same configural model to each group, **stop**. This implies that the constructs are fundamentally different for the different groups.
  - You cannot compare groups **at all** without configural invariance.

Next Steps & Partial Invariance

- #3. After configural invariance, test **metric invariance.**
  - **Equal** factor loadings across groups.
- If you don't get full metric invariance, you **can** consider **partial invariance.**
  - Allowing **some** loadings to vary across groups.
  - There are pros and cons to this.
- V & L recommend allowing individual loadings to vary **only if**:
  - **Most** loadings are invariant (you're only relaxing a few constraints).
  - It is theoretically reasonable that those particular items might measure differently in different groups.
  - You have cross-validation or replication data to support the different loadings.

More Steps

- #4. Test **scalar** invariance – **if** it's appropriate.
  - Equal item intercepts (means).
  - **Add** this constraint to your metric invariance model.
  - Meaning of this depends on underlying theory!
    - If the groups should not differ on the construct, lack of scalar invariance can signal response bias (e.g., leniency).
    - If the groups should differ on the construct, we don't *expect* scalar invariance! Differences are real, not measurement bias.
- #5. See V & L's flow chart – different study goals require different kinds of ME/I.
  - We often want to compare **latent means**.

Partial Invariance

- To reiterate: **partial invariance can be ok**.
  - As long as you are within reasonable bounds and not totally capitalizing on chance (per V & L).
- If you have partial invariance (and you know where that invariance lies), you can estimate and compare latent means.
  - You also have the option to **drop** items that don't behave well across groups.
- The flow chart in V & L is really, really handy.

IRT Methods

- We can use IRT to test for differential item functioning (DIF) across groups.
- Literally, comparing the item characteristic curves to see whether they are equivalent.
- But this is not necessarily better than MGCFA!
  - MGCFA actually makes it *easier* to tell *where* and *how* items lack invariance because we test loadings, thresholds, & uniquenesses separately.
- Up-and-coming methods:
  - Multiple indicators multiple causes (MIMIC) models
  - IRT with covariates models
- Right now, MGCFA is the standard for most applications.

Questions?

NO CLASS THURS OR LAB FRI!

For next time: Validity scales / detecting faking.

Read: Schmitt & Oswald (2006); Piedmont et al. (2000).

Reading Response: According to Piedmont et al., how are validity scales *supposed* to work? Do they?