Sensitivity of Goodness-of-Fit Indices to Lack of Measurement Invariance with Categorical Indicators and Many Groups

Boris Sokolov

LSCR HSE

16.07.2019
Research questions

- How sensitive are standard SEM goodness-of-fit indices to lack of measurement invariance with categorical data and large second-level sample sizes (10-50 groups)?
- How critical, in the same context, are different levels of non-invariance (inevitable in large samples) with respect to substantive inferences, e.g. latent means comparison?
Fit indices: CFI, TLI, RMSEA, SRMR/WRMR
Model: one factor, four items (each with four response categories).
Number of groups: \{10, 30, 50\}.
Amount of non-invariance: 9 conditions (full invariance - two with scalar non-invariance - six with scalar/metric non-invariance).
Other model misspecifications: No/one non-zero residual covariance/two non-zero residual covariances.
In sum: 81 conditions, 500 replications for each
Missing values: 10% observations in each group (MCAR)
Group sizes: 30% — 1000, 40% — 1500, 30% — 2000.
Latent means:
\( \mathcal{N}(0; 1) \)

Latent variances:
\( \mathcal{U}(0.6; 1.4) \)

Factor loadings:

1. \( \{0.75, 0.75, 0.6, 0.6\} \) in all groups
2. 1st and 2nd: trunc.  \( \mathcal{N}(0.75, 0.05; L = 0.6, U = 0.9) \)
   3rd and 4th: trunc.  \( \mathcal{N}(0.6, 0.05; L = 0.45, U = 0.75) \)
3. 1st and 2nd: trunc.  \( \mathcal{N}(0.75, 0.05; L = 0.6, U = 0.9) \)
   3rd: trunc.  \( \mathcal{N}(0.6, 0.05; L = 0.45, U = 0.75) \)
   4th:  \( \mathcal{U}(\sqrt{0.1}, 0.75) \)
4. 1st: trunc.  \( \mathcal{N}(0.75, 0.05; L = 0.6, U = 0.9) \)
   2nd:  \( \mathcal{U}(sqrt0.1, 0.9) \)
   3rd: trunc.  \( \mathcal{N}(0.6, 0.05; L = 0.45, U = 0.75) \)
   4th:  \( \mathcal{U}(\sqrt{0.1}, 0.75) \)
Thresholds:

1. **1st**: \([-0.8, 0, 0.8]\)
   - **2nd**: \([-0.8, 0, 0.8]\)
   - **3rd**: \([-0.6, 0, 0.6]\)
   - **4th**: \([-0.6, 0, 0.6]\)

2. **All**: \(\text{trunc. } \mathcal{N}(\tau_{jc}^{Cond1}, 0.05; L = \tau_{jc}^{Cond1} - 0.2, U = \tau_{jc}^{Cond1} + 0.2)\)

3. **All**: \(\text{trunc. } \mathcal{N}(\tau_{jc}^{Cond1}, 0.2; L = \tau_{jc}^{Cond1} - 0.35, U = \tau_{jc}^{Cond1} + 0.35)\)

where \(\tau_{jc}^{Cond1}\) is the threshold value for the \(j\)-th item and the \(c\)-th response category in Condition 1.
Total item variances:
\( U(0.8, 1.2) \)

Residual variances:
Item i’s total variance minus its squared loading in the fully invariant condition (0.75 or 0.6)

Residual covariances (Item 1 ~~ Item 2 and Item 3 ~~ Item 4):

1. \( \{0, 0\} \) in all groups
2. 1st: zero in all groups
   2nd: trunc. \( \mathbb{N}(0.1, 0.1, L = 0, U = 0.2) \)
3. 1st: trunc. \( \mathbb{N}(0.05, 0.1, L = -0.1, U = 0) \)
   2nd: trunc. \( \mathbb{N}(0.1, 0.1; L = 0, U = 0.2) \)
Invariance conditions

1. Full Inv.: Loadings 1 + Thresholds 1
2. Scalar 1: Loadings 1 + Thresholds 2
3. Scalar 2: Loadings 1 + Thresholds 3
4. Metric 1: Loadings 2 + Thresholds 2
5. Metric 2: Loadings 3 + Thresholds 2
6. Metric 3: Loadings 4 + Thresholds 2
7. Metric 4: Loadings 2 + Thresholds 3
8. Metric 5: Loadings 3 + Thresholds 3
9. Metric 6: Loadings 4 + Thresholds 3
Simulation and Estimation

- Simulation: R packages simsem and lavaan
- Estimation: MPLUS 7.11 (via the R package MplusAutomation)
- Estimation methods:
  - MLR
  - WLSMV (MPLUS default identification)
  - WLSMV (Wu and Estabrook’s identification approach)
Configural vs. Threshold (Wu and Estabrook)
True vs. Estimated Means Correlations

![Graph showing correlations between true and estimated means for different estimators and conditions.](image-url)
Results

- CFI seems to be the “best-performing” fit index; SRMR is the second-best (but only with MLR estimation)
- Other misspecifications negatively (and non-linearly) affects both the absolute and the relative model fit for all fit indices and invariance levels
- Second-level sample size negatively affects sample variability of fit indices but has little impact on their average values.
- Loading and intercept/threshold non-invariances generally have a multiplicative effect on model fit
- All fit indices often fail to discriminate between approximately invariant data and fully invariant data.
- It is difficult to propose universally applicable cutoff values; *ad hoc* simulations should guide researchers’ decisions.
- Even [relatively] highly non-invariant models may produce reliable (*comparable?*) latent means estimates (??)
Proposed cutoff values (MLR and WLSMV-1 estimation)

- **Configural invariance:**
  - **MLR:** $\text{CFI} > 0.985; \text{SRMR} < 0.02$
  - **WLSMV:** $\text{CFI} > 0.985$

- **Loading invariance:**
  - **MLR:** $\Delta \text{CFI} > -0.01; \Delta \text{SRMR} < 0.01$
  - **WLSMV:** $\Delta \text{CFI} > -0.005$

- **Intercept/Threshold invariance:**
  - **MLR:** $\Delta \text{CFI} > -0.01; \Delta \text{SRMR} < 0.01; \Delta \text{TLI} > -0.005; \Delta \text{RMSEA} > 0.005$
  - **WLSMV:** $\Delta \text{CFI} > -0.005; \Delta \text{TLI} > 0.00; \Delta \text{RMSEA} < 0.00$

Critical values above are based on the (approximate) average values of the 2.5th (CFI and TLI) or 97.5th (SRMR and RMSEA) percentiles of the respective fit indices averaged across all conditions in which full invariance of a given level holds.
Proposed cutoff values (WLSMV-2: Wu&Estabrook)

- Configural invariance:
  - CFI > 0.99

- Threshold Invariance:
  - $\Delta \text{CFI} > -0.005(0.002)$; $\Delta \text{CFI} > 0.00$; $\Delta \text{RMSEA} < 0.00$

- Threshold + Loading Invariance:
  - $\Delta \text{CFI} > -0.02$

Critical values above are based on the (approximate) average values of the 2.5th percentiles of the CFI averaged across all conditions in which full invariance of a given level holds.
Thank you very much for your attention!

Please send your questions, comments and feedback at bssokolov@gmail.com


