Measurement Invariance Workshop Fall 2018

Michael Broda, Ph.D.

VCU School of Education

9/6/2018

Agenda

• Brief Overview of Measurement Invariance (MI)
• Conducting MI Analysis in R with the lavaan package
• Stepwise Analysis & Recommendations
• Practice and Take-Aways

Background - Measurement Invariance (1)

• When we administer a scale, instrument, etc., we make many assumptions about the underlying measurement properties
• Key among them - the instrument functions the same for all participants
• Items are interpreted equivalently, have same meaning, etc.
• This is measurement invariance - invariance is a good thing in this case!

Background - Measurement Invariance (2)

• This is a statistically testable assumption using the tools of SEM / CFA
• We may want to test MI for several reasons:
• Our sample is heterogeneous, and we want to show that the scale we are using functions equivalently between groups so we can make an accurate comparison
• Or, maybe we expect that two groups are interpreting a scale differently, and that different interpretations are novel/useful, so we want to show they are different

Let’s Use Some Sample Data

• Dataset is s4s_simulated.dta (Stata format)
• Simulated from VCU Spit for Science Registry, College Behavioral and Emotional Health Institute
• http://cobe.vcu.edu
• 1,785 students who took the brief version of the Almost Perfect Scale (Revised)
• (Slaney, Mobley, Trippi, Ashby, & Johnson, 1996)
• Two dimensions of perfectionism: high standards and discrepancy (self-criticality)
Load in and Summarize the Data

```r
library(haven)
s4s <- read_dta("s4s_simulated.dta")
psych::describe(s4s, fast = TRUE)
```

<table>
<thead>
<tr>
<th>vars</th>
<th>n</th>
<th>mean</th>
<th>sd</th>
<th>min</th>
<th>max</th>
<th>range</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>hs1</td>
<td>1785</td>
<td>0.00</td>
<td>1.0</td>
<td>-4.37</td>
<td>3.29</td>
<td>7.66</td>
<td>0.02</td>
</tr>
<tr>
<td>hs2</td>
<td>1785</td>
<td>0.00</td>
<td>1.0</td>
<td>-3.47</td>
<td>3.60</td>
<td>7.08</td>
<td>0.02</td>
</tr>
<tr>
<td>hs3</td>
<td>1785</td>
<td>0.00</td>
<td>1.0</td>
<td>-3.07</td>
<td>3.11</td>
<td>6.18</td>
<td>0.02</td>
</tr>
<tr>
<td>dis1</td>
<td>1785</td>
<td>0.00</td>
<td>1.0</td>
<td>-3.26</td>
<td>2.95</td>
<td>6.21</td>
<td>0.02</td>
</tr>
<tr>
<td>dis2</td>
<td>1785</td>
<td>0.00</td>
<td>1.0</td>
<td>-3.10</td>
<td>3.58</td>
<td>6.68</td>
<td>0.02</td>
</tr>
<tr>
<td>dis3</td>
<td>1785</td>
<td>0.00</td>
<td>1.0</td>
<td>-3.82</td>
<td>2.88</td>
<td>6.70</td>
<td>0.02</td>
</tr>
<tr>
<td>scimajor</td>
<td>1785</td>
<td>0.51</td>
<td>0.5</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.01</td>
</tr>
<tr>
<td>firstgen</td>
<td>1785</td>
<td>0.50</td>
<td>0.5</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Create a CFA Model and Run it in lavaan

```r
mdp.model <- ' discrepancy =~ dis1 + dis2 + dis3
  high standards =~ hs1 + hs2 + hs3 '
library(lavaan)
fit <- cfa(mdp.model, data=s4s)
summary(fit, fit.measures=TRUE)
```

```
## lavaan 0.6-2 ended normally after 22 iterations
##
## Optimization method      NLMINB
## Number of free parameters 13
##
## Number of observations    1785
##
## Estimator                ML
## Model Fit Test Statistic 56.485
## Degrees of freedom       8
## P-value (Chi-square)      0.000
##
## Model test baseline model:
##
## Minimum Function Test Statistic 3676.180
## Degrees of freedom          15
## P-value                      0.000
##
## User model versus baseline model:
##
## Comparative Fit Index (CFI) 0.987
## Tucker-Lewis Index (TLI)     0.975
```
## Loglikelihood and Information Criteria:

<table>
<thead>
<tr>
<th>Model</th>
<th>Loglikelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>User model (H0)</td>
<td>-13383.983</td>
</tr>
<tr>
<td>Unrestricted model (H1)</td>
<td>-13355.741</td>
</tr>
</tbody>
</table>

## Number of free parameters 13

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akaike (AIC)</td>
<td>26793.967</td>
</tr>
<tr>
<td>Bayesian (BIC)</td>
<td>26865.300</td>
</tr>
<tr>
<td>Sample-size adjusted Bayesian (BIC)</td>
<td>26824.000</td>
</tr>
</tbody>
</table>

## Root Mean Square Error of Approximation:

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSEA</td>
<td>0.058</td>
</tr>
<tr>
<td>90 Percent Confidence Interval</td>
<td>0.045 0.073</td>
</tr>
<tr>
<td>P-value RMSEA &lt;= 0.05</td>
<td>0.154</td>
</tr>
</tbody>
</table>

## Standardized Root Mean Square Residual:

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRMR</td>
<td>0.036</td>
</tr>
</tbody>
</table>

## Parameter Estimates:

### Information

<table>
<thead>
<tr>
<th>Information saturated (h1) model</th>
<th>Structured</th>
</tr>
</thead>
</table>

### Standard Errors

<table>
<thead>
<tr>
<th>Standard Errors</th>
<th>Standard</th>
</tr>
</thead>
</table>

## Latent Variables:

| Discrepancy =~ | Estimate | Std.Err | z-value | P(>|z|) |
|----------------|----------|---------|---------|--------|
| dis1           | 1.000    |         |         |        |
| dis2           | 1.145    | 0.043   | 26.445  | 0.000  |
| dis3           | 1.228    | 0.047   | 26.285  | 0.000  |

| Highstandards =~ | Estimate | Std.Err | z-value | P(>|z|) |
|------------------|----------|---------|---------|--------|
| hs1               | 1.000    |         |         |        |
| hs2               | 1.042    | 0.038   | 27.615  | 0.000  |
| hs3               | 0.994    | 0.036   | 27.368  | 0.000  |

## Covariances:

| Discrepancy ~~ Highstandards | Estimate | Std.Err | z-value | P(>|z|) |
|------------------------------|----------|---------|---------|--------|
| discrepancy ~ highstandards  | -0.022   | 0.015   | -1.471  | 0.141  |

## Variances:

| Discrepancy                  | Estimate | Std.Err | z-value | P(>|z|) |
|-----------------------------|----------|---------|---------|--------|
| .dis1                        | 0.535    | 0.022   | 23.873  | 0.000  |
| .dis2                        | 0.390    | 0.022   | 17.758  | 0.000  |
| .dis3                        | 0.299    | 0.023   | 13.211  | 0.000  |
| .hs1                         | 0.424    | 0.022   | 19.546  | 0.000  |
| .hs2                         | 0.374    | 0.022   | 17.170  | 0.000  |
| .hs3                         | 0.430    | 0.022   | 19.855  | 0.000  |
CFA Results for One Group (All Participants)

<table>
<thead>
<tr>
<th>term</th>
<th>estimate</th>
<th>std.error</th>
<th>statistic</th>
<th>p.value</th>
<th>std.all</th>
</tr>
</thead>
<tbody>
<tr>
<td>discrepancy</td>
<td>1.0000000</td>
<td>0.0000000</td>
<td>NA</td>
<td>NA</td>
<td>0.6819612</td>
</tr>
<tr>
<td>discrepancy</td>
<td>1.1452224</td>
<td>0.0433053</td>
<td>26.445325</td>
<td>0.0000000</td>
<td>0.7809973</td>
</tr>
<tr>
<td>discrepancy</td>
<td>1.2277465</td>
<td>0.0467094</td>
<td>26.284776</td>
<td>0.0000000</td>
<td>0.8372754</td>
</tr>
<tr>
<td>highstandards</td>
<td>1.0000000</td>
<td>0.0000000</td>
<td>NA</td>
<td>NA</td>
<td>0.7589999</td>
</tr>
<tr>
<td>highstandards</td>
<td>1.0424683</td>
<td>0.0377500</td>
<td>27.615046</td>
<td>0.0000000</td>
<td>0.7912334</td>
</tr>
<tr>
<td>highstandards</td>
<td>0.9940826</td>
<td>0.0363227</td>
<td>27.368104</td>
<td>0.0000000</td>
<td>0.7545086</td>
</tr>
</tbody>
</table>

CFA Results for Two-Factor Model
<table>
<thead>
<tr>
<th></th>
<th>dis1</th>
<th>dis1</th>
<th>dis2</th>
<th>dis2</th>
<th>dis3</th>
<th>dis3</th>
<th>hs1</th>
<th>hs1</th>
<th>hs2</th>
<th>hs2</th>
<th>hs3</th>
<th>hs3</th>
<th>discrepancy</th>
<th>discrepancy</th>
<th>discrepancy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5346292</td>
<td>0.0223952</td>
<td>23.872511</td>
<td>0.0000000</td>
<td>0.5349289</td>
<td></td>
<td>0.3898248</td>
<td>0.0219516</td>
<td>17.758414</td>
<td>0.0000000</td>
<td>0.3900433</td>
<td></td>
<td>0.2988024</td>
<td>0.0226177</td>
<td>13.210971</td>
</tr>
<tr>
<td></td>
<td>0.4236816</td>
<td>0.0216758</td>
<td>19.546294</td>
<td>0.0000000</td>
<td>0.4239191</td>
<td></td>
<td>0.3737403</td>
<td>0.0217670</td>
<td>17.170049</td>
<td>0.0000000</td>
<td>0.3739498</td>
<td></td>
<td>0.4304754</td>
<td>0.0216805</td>
<td>19.855381</td>
</tr>
<tr>
<td></td>
<td>0.4648105</td>
<td>0.0313086</td>
<td>14.846121</td>
<td>0.0000000</td>
<td>1.0000000</td>
<td></td>
<td>0.5757581</td>
<td>0.0344497</td>
<td>16.713030</td>
<td>0.0000000</td>
<td>1.0000000</td>
<td></td>
<td>-0.0149601</td>
<td>-1.471001</td>
<td>0.1412907</td>
</tr>
<tr>
<td></td>
<td>0.0220063</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0425392</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Good So Far- But How Do You Test Measurement Invariance?**

- There are multiple levels of measurement invariance
- Configural (Baseline)
- Metric (Weak)
- Scalar (Strong)
- Error Variance (Strict)
- And more! But these four are the most commonly tested.
- They proceed in steps: if your data meets the criteria, you test more restrictive levels of invariance
- In general, scalar invariance is recommended to allow for comparisons between groups

**Configural Invariance**

- First step in testing MI - this is the “baseline” model
- Is the configuration of items and latent variables consistent between groups?
- We will use a variable, scimajor, to define our groups
- If scimajor = 1, then student is a science major
- If scimajor = 0, then student is not a science major
- Does the APS-R function equivalently for students in both groups?

**Configural Invariance Visual**

Science Majors
Non-Science Majors

**Metric Invariance**
- Second step in testing MI - compared to the configural model
- Here, we constrain the item loadings (slopes) to be equal between the groups
- Are item loadings (slopes) consistent between groups?
- Sometimes called “weak” invariance (it’s the first one you test)
Metric Invariance Visual

Science Majors

Non-Science Majors

Scalar Invariance
- Third step in testing MI - compared to the metric model
- Here, we constrain the intercepts (thresholds) to be equal between the groups
- Are item intercepts (thresholds) consistent between groups?
- Also called “strong” invariance
Scalar Invariance Visual

Science Majors

Non-Science Majors

Error Variance Invariance

- Fourth step in testing MI - compared to the scalar model
- Here, we constrain the item-level error variances to be equal between the groups
- Are item error variances consistent between groups?
- Also called "strict" invariance
Error Variance Invariance Visual

Science Majors

Non-Science Majors

Quick Summary of Invariance Types
<table>
<thead>
<tr>
<th>Invariance Type</th>
<th>What is Constrained?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configural (Baseline)</td>
<td>Arrangement of Items and Factors</td>
</tr>
<tr>
<td>Metric (Weak)</td>
<td>Item Loadings (Slopes)</td>
</tr>
<tr>
<td>Scalar (Strong)</td>
<td>Item Intercepts</td>
</tr>
<tr>
<td>Error Variance (Strict)</td>
<td>Item-Specific Error Terms</td>
</tr>
</tbody>
</table>

The more levels of invariance you can establish, the stronger the case that your measure does not function differently between groups.

**Assessing Measurement Invariance in lavaan**

```r
library(lavaan)
library(semTools)

##
## This is semTools 0.5-0
## All users of R (or SEM) are invited to submit functions or ideas for functions.

mdp.model <- 'discrepancy =~ dis1 + dis2 + dis3
                  high standards =~ hs1 + hs2 + hs3'
measurementInvariance(model = mdp.model,
                         data = s4s,
                         group = "scimajor",
                         strict = TRUE)

## Measurement invariance models:
##
## Model 1: fit.configural
## Model 2: fit.loadings
## Model 3: fit.intercepts
## Model 4: fit.residuals
## Model 5: fit.means
##
## Chi Square Difference Test
##
## |               | Df | AIC  | BIC  | Chisq | Chisq diff | Df diff | Pr(>Chisq) |
##|---------------|----|------|------|-------|------------|---------|------------|
##| fit.configural| 16 | 26833| 27042| 72.094|            |         |            |
##| fit.loadings  | 20 | 26831| 27017| 77.313| 5.2187     | 4       | 0.2656     |
##| fit.intercepts| 24 | 26826| 26990| 80.180| 2.8669     | 4       | 0.5803     |
##| fit.residuals | 30 | 26815| 26946| 81.338| 1.1584     | 6       | 0.9789     |
```
What to Look For in the Results of MI Testing (1)

- The clearest evidence for MI testing is the Chi Square ($\chi^2$) Difference Test
- Each model is tested against the previous type
- Configural (the baseline) vs. Metric
- Metric vs. Scalar
- Scalar vs. Error Variance
- The difference in $\chi^2$ is taken for each comparison, and then using the difference in degrees of freedom, we are able to get a $p$-value
- A non-significant $p$-value is a good thing here - it means that the model “passes the test” and meets the higher level of invariance
- A significant $p$-value tells you the opposite - the model “fails” and does not meet the higher level of invariance

Let’s Have a Look at the Output

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>Chi sq</th>
<th>Chi sq diff</th>
<th>Df diff</th>
<th>Pr(&gt;Chi sq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>fit.configural</td>
<td>16</td>
<td>26833</td>
<td>27042</td>
<td>72.094</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fit.loadings</td>
<td>20</td>
<td>26831</td>
<td>27017</td>
<td>77.313</td>
<td>5.2187</td>
<td>4</td>
<td>0.2656</td>
</tr>
<tr>
<td>fit.intercepts</td>
<td>24</td>
<td>26826</td>
<td>26990</td>
<td>80.180</td>
<td>2.8669</td>
<td>4</td>
<td>0.5803</td>
</tr>
<tr>
<td>fit.residuals</td>
<td>30</td>
<td>26815</td>
<td>26946</td>
<td>81.338</td>
<td>1.1584</td>
<td>6</td>
<td>0.9789</td>
</tr>
<tr>
<td>fit.means</td>
<td>32</td>
<td>26811</td>
<td>26931</td>
<td>81.370</td>
<td>0.0319</td>
<td>2</td>
<td>0.9842</td>
</tr>
</tbody>
</table>

- This looks all good! All $p$-values are non-significant, so we have met even the more stringent tests of invariance
- We can say with confidence that this scale does not function differently for science majors vs. non-science majors

Here’s the Tricky Part...

- Unfortunately, it isn’t always this easy!
- In large samples, $\chi^2$ Difference Tests are not the best judge of MI
- They are often too conservative (they show that a model fails when it actually doesn’t)
- So what else can we use to make a decision?
• Confirmatory Fit Index (CFI)
• Root Mean Square Error of Approximation (RMSEA)
• Just like we did with $\chi^2$, we can look at how these values change when we test each level of MI
• \textbf{Lower} is better for RMSEA, and \textbf{higher} is better for CFI

\textbf{Back to the R Output!}

-At the very bottom of the output from our measurement invariance model, we get a summary of the changes in CFI and RMSEA:

<table>
<thead>
<tr>
<th>Fit measures</th>
<th>cfi</th>
<th>rmsea</th>
<th>cfi.delta</th>
<th>rmsea.delta</th>
</tr>
</thead>
<tbody>
<tr>
<td>fit.configural</td>
<td>0.985</td>
<td>0.063</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>fit.loadings</td>
<td>0.984</td>
<td>0.057</td>
<td>0.000</td>
<td>0.006</td>
</tr>
<tr>
<td>fit.intercepts</td>
<td>0.985</td>
<td>0.051</td>
<td>0.000</td>
<td>0.005</td>
</tr>
<tr>
<td>fit.residuals</td>
<td>0.986</td>
<td>0.044</td>
<td>0.001</td>
<td>0.007</td>
</tr>
<tr>
<td>fit.means</td>
<td>0.987</td>
<td>0.042</td>
<td>0.001</td>
<td>0.002</td>
</tr>
</tbody>
</table>

-In this case, we already know that our model passes the MI test, but let's check the RMSEA and CFI anyway... - We get four columns: cfi, rmsea, cfi.delta, and rmsea.delta - The “deltas” tell us how much the fit statistics changed for each test - CFI went up (.001) from scalar to error variance, RMSEA went down

\textbf{How Much Change Are We Looking For?}

• Conducted extensive simulation studies to test different cutoffs for fit values
• A CFI \textbf{decrease} $\geq .01$, along with an RMSEA \textbf{increase} $\geq .015$ in RMSEA, would indicate \textbf{noninvariance} (bad thing).
• Big Idea: you want CFI to stay the same, and hopefully not get \textit{smaller}, and you want RMSEA to stay the same, and hopefully not get \textit{larger}
• If your model meets these requirements, you can usually still claim MI even if you have a significant $\chi^2$ Difference Test.

\textbf{How About a Little Practice?}

• We have another grouping variable, firstgen, to define our groups
• If firstgen = 1, then student is a first-generation college attendee
• If firstgen = 0, then student is a continuing-generation college attendee
• Key Question:
  \textit{Does the APS-R function equivalently for students in both groups?}
• Use R to test for measurement invariance
• You can use the R code from the presentation to conduct the MI analysis
• You will just need to change the variable sci_major to firstgen
Wrap Up

- Hopefully Measurement Invariance isn’t so scary!
- The trickiest part is knowing what to look for and how to make a judgment
- There also is such a thing as partial measurement invariance
- You can constrain everything between groups except for one parameter that is giving you trouble
- This gets a lot trickier! You need to not only know how to free the parameter, you also need to justify why that item doesn’t work consistently and interpret accordingly

References/ Recommended Readings


Contact Information

Michael Broda, Ph.D.
Assistant Professor
Department of Foundations of Education
VCU School of Education
mdbroda@vcu.edu