Measurement Invariance in Longitudinal Data

Ji Hoon Ryoo, Ph.D.
Postdoctoral Fellow
National Center for Research on Rural Education (R²ED)
Overview

• Introduction to Measurement Invariance in Longitudinal Data
• Measurement Tools
• Background on Item Response Theory (IRT)
• Dichotomous and Polytomous IRT Models
• Longitudinal Invariance
• Linking Repeated Measures with Longitudinal Invariance Items
• Model Selection in Longitudinal Data

• Data: PTRS in Getting Ready project
• Invariance Results
• Linking by Common Items with Multiple Groups
• Model Selection on Longitudinal Latent Variables
• Conclusion
Intro to Longitudinal Invariance

• The examination of longitudinal measurement invariance can be used to determine whether items on a particular instrument assess the same attribute across time (Horn & McArdle, 1992; Meredith, 1993)

• If a scale used to assess a particular attribute does not exhibit evidence of longitudinal invariance, then the interpretation of changes in mean scores and correlations between time points may be ambiguous (Horn & McArdle, 1992)

• Although researchers often implicitly assume that administering the same instrument across multiple time points ensures that the same attribute is being assessed (i.e., longitudinal invariance), this empirical hypothesis is rarely tested, and when it is, it is often rejected (de Frias & Dixon, 2005; Maitland, Dixon, Hultsch, & Hertzog, 2001; Motl, Dishman, Birnbaum, & Lytle, 2005)
Intro to Longitudinal Invariance

• How to assess measurement invariance?
  – Does the instrument measure the same construct(s) over time?
  – Does each question measure the construct to the same degree?

• What if measurement is not invariant over time?
  – It is not possible to develop an instrument that is invariant universally.
  – Every instrument may or may not include problematic items.

• How is ability (or attitude) estimated if some portion of the measurement model is varying?
  – Should we avoid using problematic items to estimate the ability (or attitude)?
  – Can we use the ability (or attitude) scores estimated by subtest consisting of longitudinal invariant items?
Measurement Tools

• Classical test theory (CTT)
  – Linear model
  – Weak assumptions
  – Item and person parameters are sample dependent

• Item response theory (IRT)
  – Nonlinear model
  – Strong assumptions
  – Item and person parameters are sample independent if model fits the test data

• Why does IRT better fit to test the longitudinal invariance?
  – Short version is allowed even if it is not already validated
  – Both item and person fit statistics are provided
Background on IRT

- IRT consists of a set of latent variable models for responses to test or questionnaire items

- IRT models are divided into two groups on the basis of how items are scored – dichotomous vs. polytomous

- Most models in both cases are unidimensional: Item responses depend on a single latent variable that explains all the statistical associations among the item scores

- The mathematical relation between the person’s score on the latent variable and his or her item response is described by the item response function
Dichotomous IRT models

• The response probabilities for the three-parameter logistic model (3PL; Birnbaum 1968): let $X_{ij}$ be the response of examinee $i$ to item $j$, then

$$P_{ij} = P(X_{ij} = 1 \mid \theta_i, a_j, b_j, c_j)$$

$$= c_j + (1 - c_j) \frac{\exp[D a_j (\theta_i - b_j)]}{1 + \exp[D a_j (\theta_i - b_j)]}$$

Equation (1)

where $\theta_i$ is an examinee’s ability on a single construct, $a_j$ is the item discrimination, $b_j$ is the item difficulty, $c_j$ is the lower asymptote (guessing parameter), and $D$ is a scaling constant.

• If the guessing parameter is constrained to be zero, Equation 1 becomes the two-parameter logistic model (2PL; Birnbaum 1968)

• If it is further constrained so that the discrimination parameters for all items are equal, it becomes the one-parameter logistic model (1PL)
Item Characteristic Curves for 3PL and 2PL

- 3PL
- vs.
- 2PL

![Item Characteristic Curves](image_url)
**Item & Test Information Curves**

- Item information curves vs. Test information curves
Polytomous IRT Models (1)

The graded response model (GRM; Samejima 1969)

\[ P_{ijk} = \tilde{P}_{ijk} - \tilde{P}_{ij(k+1)} \]

where \( \tilde{P}_{ijk} = \tilde{P}(X_{ij} = k \mid \theta_i, a_j, b_{jk}) \)

\[ = \begin{cases} 
1, & \text{if } k = 1 \\
\exp[D a_j (\theta_i - b_{jk})] \\
1 + \exp[D a_j (\theta_i - b_{jk})], & \text{if } 2 \leq k \leq K_j \\
0, & \text{if } k > K_j 
\end{cases} \]

Models

- Constrained version: \( a_i = a \) for all items
- Unconstrained version: different \( a_i \) per item
**Polytomous IRT Models (2)**

- The generalized partial credit model (GPCM; Muraki 1992)

\[
P_{ijk} = P(X_{ij} = v | \theta_i, a_j, b_{jk})
\]

\[
= \frac{\exp\left[\sum_{v=1}^{k} Da_j (\theta_i - b_{jv})\right]}{\sum_{h=1}^{K_i} \exp\left[\sum_{v=1}^{h} Da_j (\theta_i - b_{jv})\right]}
\]

- Models
  - Constrained version: \( a_i = a \) for all items
  - Unconstrained version: different \( a_i \) per item
Item Characteristic Curve (ICC)

- ICCs (Unconstrained)
**Item Characteristic Curve (ICC)**

- ICCs (Constrained)

---

**Item Response Category Characteristic Curves**

*Item: Item 1*

- Probability
- Latent Trait

**Item Response Category Characteristic Curves**

*Item: Item 3*

- Probability
- Latent Trait
Item and Test Information Curves

- Test consisting of 5 items whose number of categories in each item is 5
**Longitudinal Invariance (1)**

- **Step 1:**
  - Configural invariance: test the invariance of the dimensionality
    - Pairwise association (Non-significant $\rightarrow$ Independent item)
    - Eigenvalues (Scree plot; Elbow rule)
    - Cronbach’s alpha ($> 0.80$)

- What if the dimensionality changes over time?
  - The true developmental changes in the attribute may be confounded with changes in other variables such as item functioning of the instrument
  - The factor scores are unclear and do not represent the quantification of the construct at each time point
Longitudinal Invariance (2)

• Step 2:
  • Longitudinal metric invariance: examining whether individual items display equivalent parameter estimates on the latent construct across several developmental assessments
    • Fitting an IRT model and comparing model fit by the likelihood ratio test (LRT) or Delta method
    • Parameter estimates (with 5% error rate)
  • From a psychometric perspective, this implies that the factor loadings of individual items indexing the construct do not vary with development
Linking repeated measures with longitudinal invariant items

- By investigating configural and metric invariance, we can construct the subset of the test whose items are all longitudinally invariant.

- Next we link all scales onto one scale with the longitudinally invariant items. Thus, we can compare scaled scores over time.

\[ \theta_T = A \cdot \theta_F + B = \theta_F^* \]

- Stocking-Lord’s Linking: Estimate the linking constants by minimizing the sum of squared differences between test information curves for the common items.

- Haebara’s Linking: Estimate the linking constants by minimizing the sum of squared differences between item characteristic curves for the common items.
Model Selection in Longitudinal Data

• After obtaining latent scores (scaled scores linked by common items), we fit statistical models to longitudinal data consisting of latent scores

• Common statistical models
  – Polynomial models
  – Fractional polynomial models
  – Trigonometric models
  – Spline (piecewise) models

• Two packages, ltm (Rizopoulos, 2006) and plink (Weeks, 2010), in R program were used
**Data: Parent-Teacher Relationship Scale**

- Parent-Teacher Relationship Scale (PTRS; Vickers & Minke, 1995)
  - 24 items
  - 5-point Likert scale
  - Assessing the cohesion and adaptability of parent-teacher relationship system via two subscales, joining and communication to others
  - Data collected in the Getting Ready project (CYFS)
  - Data consisting of 4 repeated measures completed by parents and teachers

<table>
<thead>
<tr>
<th>Subscales</th>
<th>Parent</th>
<th>Teacher</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joining (items 1 to 19)</td>
<td>Time 1 to Time 4</td>
<td>Time 1 to Time 4</td>
</tr>
<tr>
<td>Communication to others (items 20 to 24)</td>
<td>Time 1 to Time 4</td>
<td>Time 1 to Time 4</td>
</tr>
</tbody>
</table>
# PTRS (Parent) - Joining subscale

- Configural and longitudinal metric invariance results

<table>
<thead>
<tr>
<th></th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 3</th>
<th>Time 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size (# of items)</td>
<td>183 (17)</td>
<td>159 (8)</td>
<td>119 (11)</td>
<td>87 (5)</td>
</tr>
<tr>
<td><strong>Configural Invariance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scree Plot (1st eigenvalue)</td>
<td>One (6.840)</td>
<td>One (3.892)</td>
<td>One (3.781)</td>
<td>One (2.756)</td>
</tr>
<tr>
<td>Pairwise Association (items)</td>
<td>Items 5 &amp; 7</td>
<td>Items 15, 9, 14, 11, 8, 12, &amp; 19</td>
<td>Items 19, 2, 8, 11, 12, &amp; 4</td>
<td>Items 11, 4, 8, 11, 12, &amp; 14</td>
</tr>
<tr>
<td>Cronbach's $\alpha$</td>
<td>0.8989</td>
<td>0.8326</td>
<td>0.8811</td>
<td>0.8989</td>
</tr>
<tr>
<td>No response on a category</td>
<td>Item 5</td>
<td>Items 5, 6, 7, &amp; 10</td>
<td>Items 5, 6, 12, &amp; 13</td>
<td>Items 3, 5, 6, 7, 12, 13, 15, 17, 18, &amp; 19</td>
</tr>
<tr>
<td><strong>Metric Invariance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRT model (equivalent)</td>
<td>GRM2 (GRM1)</td>
<td>GRM2 (GRM1)</td>
<td>GRM2 (GRM1)</td>
<td>GRM1 (GRM2)</td>
</tr>
</tbody>
</table>
**PTRS (Teacher) - Joining subscale**

- Configural and longitudinal metric invariance results

<table>
<thead>
<tr>
<th></th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 3</th>
<th>Time 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size (# of items)</td>
<td>171 (12)</td>
<td>184 (10)</td>
<td>131 (11)</td>
<td>104 (13)</td>
</tr>
<tr>
<td><strong>Configural Invariance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scree Plot (1st eigenvalue)</td>
<td>One (6.106)</td>
<td>One (5.522)</td>
<td>One (6.124)</td>
<td>One (7.943)</td>
</tr>
<tr>
<td>Pairwise Association (items)</td>
<td>Item 19</td>
<td>No item</td>
<td>Items 14 &amp; 2</td>
<td>Items 19 &amp; 14</td>
</tr>
<tr>
<td>Cronbach's $\alpha$</td>
<td>0.9053</td>
<td>0.9053</td>
<td>0.9053</td>
<td>0.9414</td>
</tr>
<tr>
<td>No response on a category</td>
<td>Items 1, 3, 5, 7, 13, &amp; 17</td>
<td>Items 1, 2, 3, 6, 7, 9, 11, 13, &amp; 19</td>
<td>Items 1, 3, 6, 7, 13, &amp; 19</td>
<td>Items 1, 2, 3, &amp; 9</td>
</tr>
<tr>
<td><strong>Metric Invariance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRT model (equivalent)</td>
<td>GRM2 (GRM1)</td>
<td>GRM2 (GRM1)</td>
<td>GRM2 (GRM1)</td>
<td>GRM2 (GRM1)</td>
</tr>
</tbody>
</table>
# PTRS (Parent) - Comm. to others subscale

- Configural and longitudinal metric invariance results

<table>
<thead>
<tr>
<th></th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 3</th>
<th>Time 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size (# of items)</td>
<td>200 (5)</td>
<td>169 (4)</td>
<td>125 (5)</td>
<td>90 (5)</td>
</tr>
<tr>
<td><strong>Configural Invariance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scree Plot (1st eigenvalue)</td>
<td>One (3.884)</td>
<td>One (3.061)</td>
<td>One (3.414)</td>
<td>One (3.563)</td>
</tr>
<tr>
<td>Pairwise Association (items)</td>
<td>No item</td>
<td>No item</td>
<td>No item</td>
<td>No item</td>
</tr>
<tr>
<td>Cronbach's α</td>
<td>0.9274</td>
<td>0.8939</td>
<td>0.8811</td>
<td>0.8989</td>
</tr>
<tr>
<td>No response on a category</td>
<td>No item</td>
<td>Item 23</td>
<td>No item</td>
<td>No item</td>
</tr>
<tr>
<td><strong>Metric Invariance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRT model (equivalent)</td>
<td>GRM1</td>
<td>GRM1 (GPCM2)</td>
<td>GRM1</td>
<td>GRM1</td>
</tr>
</tbody>
</table>
**PTRS (Teacher)**
- **Comm. to others subscale**

- Configural and longitudinal metric invariance results

<table>
<thead>
<tr>
<th></th>
<th>Time 1</th>
<th>Time 2</th>
<th>Time 3</th>
<th>Time 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size (# of items)</td>
<td>197 (5)</td>
<td>192 (3)</td>
<td>137 (4)</td>
<td>107 (4)</td>
</tr>
<tr>
<td><strong>Configural Invariance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scree Plot (1st eigenvalue)</td>
<td>One (3.465)</td>
<td>One (2.071)</td>
<td>One (2.973)</td>
<td>One (2.644)</td>
</tr>
<tr>
<td>Pairwise Association (items)</td>
<td>No item</td>
<td>No item</td>
<td>No item</td>
<td>No item</td>
</tr>
<tr>
<td>Cronbach's $\alpha$</td>
<td>0.8842</td>
<td>0.7622</td>
<td>0.8278</td>
<td>0.8217</td>
</tr>
<tr>
<td>No response on a category</td>
<td>No item</td>
<td>Items 20 &amp; 23</td>
<td>Item 21</td>
<td>Item 20</td>
</tr>
<tr>
<td><strong>Metric Invariance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRT model (equivalent)</td>
<td>GRM1 (GRM2)</td>
<td>GRM2 (GPCM2)</td>
<td>GRM1 (GRM2)</td>
<td>GRM 2</td>
</tr>
</tbody>
</table>
Results on Longitudinal Metric Invariance

- List of longitudinal invariant items and selected models for subscales

<table>
<thead>
<tr>
<th>Longitudinal invariant items</th>
<th>Parent</th>
<th>Teacher</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joining</td>
<td>Joining</td>
<td>Communication to others</td>
</tr>
<tr>
<td>2 out of 19 items (10.5%)</td>
<td>4 out of 5 items (80%)</td>
<td>7 out of 19 items (36.8%)</td>
</tr>
<tr>
<td>4 out of 5 items (80%)</td>
<td>1 out of 5 items (20%)</td>
<td></td>
</tr>
<tr>
<td>7 out of 19 items (36.8%)</td>
<td>20, 21, 22, and 24</td>
<td></td>
</tr>
</tbody>
</table>

- Extreme measurements and common items used in linking process

<table>
<thead>
<tr>
<th>Common items</th>
<th>Parent - Joining</th>
<th>Parent - Comm. to others</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 and 16</td>
<td></td>
<td>20, 21, 22, and 24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Percent of total information provided by longitudinal invariant items</th>
<th>Time 1 (13.87%)</th>
<th>Time 1 (78.86%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time 2 (30.21%)</td>
<td>Time 2 (100%)</td>
<td></td>
</tr>
<tr>
<td>Time 3 (21.79%)</td>
<td>Time 3 (79.72%)</td>
<td></td>
</tr>
<tr>
<td>Time 4 (39.93%)</td>
<td>Time 4 (82.38%)</td>
<td></td>
</tr>
</tbody>
</table>
## Results on Linking

- Linking constants for parent joining subscale

<table>
<thead>
<tr>
<th></th>
<th>Time 2 to Time 1</th>
<th></th>
<th>Time 3 to Time 1</th>
<th></th>
<th>Time 4 to Time 1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Haebara</td>
<td>0.274</td>
<td>-0.567</td>
<td>0.543</td>
<td>-0.018</td>
<td>0.565</td>
<td>-0.115</td>
</tr>
<tr>
<td>Stocking-Lord</td>
<td>0.381</td>
<td>-0.258</td>
<td>0.532</td>
<td>-0.054</td>
<td>0.657</td>
<td>0.032</td>
</tr>
</tbody>
</table>

- Linking constants for parent communication to others subscale

<table>
<thead>
<tr>
<th></th>
<th>Time 2 to Time 1</th>
<th></th>
<th>Time 3 to Time 1</th>
<th></th>
<th>Time 4 to Time 1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Haebara</td>
<td>0.705</td>
<td>-0.009</td>
<td>0.661</td>
<td>-0.088</td>
<td>0.454</td>
<td>-0.252</td>
</tr>
<tr>
<td>Stocking-Lord</td>
<td>0.810</td>
<td>0.181</td>
<td>0.755</td>
<td>0.073</td>
<td>0.585</td>
<td>-0.042</td>
</tr>
</tbody>
</table>
Change over time on joining subscale

- Box plots and connected means with lowess
Change over time on communication to others subscale

- Box plots and connected means with lowess
Model Selection on Longitudinal Latent Variables

- Fit linear mixed model to longitudinal data
  - Joining subscale
  - Factor score
    - fitted by a fractional polynomial (resulting in a log model)
    - the treatment is not significant
  - Subscale
    - fitted by a polynomial model (resulting in a mean model)
    - the treatment is significant

<table>
<thead>
<tr>
<th></th>
<th>AIC</th>
<th>BIC</th>
<th>$\chi^2$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Factor</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>606.88</td>
<td>618.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log</td>
<td>592.88</td>
<td>607.85</td>
<td>16.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Treatment</td>
<td>591.67</td>
<td>610.39</td>
<td>3.20</td>
<td>0.07</td>
</tr>
<tr>
<td><strong>Subscale</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>469.15</td>
<td>480.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>470.18</td>
<td>488.89</td>
<td>2.97</td>
<td>0.23</td>
</tr>
<tr>
<td>Treatment</td>
<td>466.25</td>
<td>481.22</td>
<td>4.90</td>
<td>0.03</td>
</tr>
</tbody>
</table>
**Conclusion**

- The criteria used in testing measurement invariance in longitudinal data are somewhat subjective.
- It is crucial to test measurement invariance when the goal is to articulate change in a latent construct over time. In this talk, I tried to provide a unified framework for constructing measurement invariance in longitudinal data.
- The purpose of this talk was to provide a demonstrate of methods for evaluating longitudinal invariance, not to make conclusions about the PTRS or Getting Ready Project. Because of attrition, and consequently a reduced sample size at times 3 and 4, statistical inferences are likely biased.
References


Thank you!

For more information, please contact:
Ji Hoon Ryoo
Phone #: (402) 472-6190
E-mail: jryoo2@unl.edu