CYFS nebraska center for research
children youth families & schools

research methodology series

Measurement Invariance in Longitudinal Data

Ji Hoon Ryoo, Ph.D.

Postdoctoral Fellow

National Center for Research on Rural Education (R^2ED)

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

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

Equation (1)

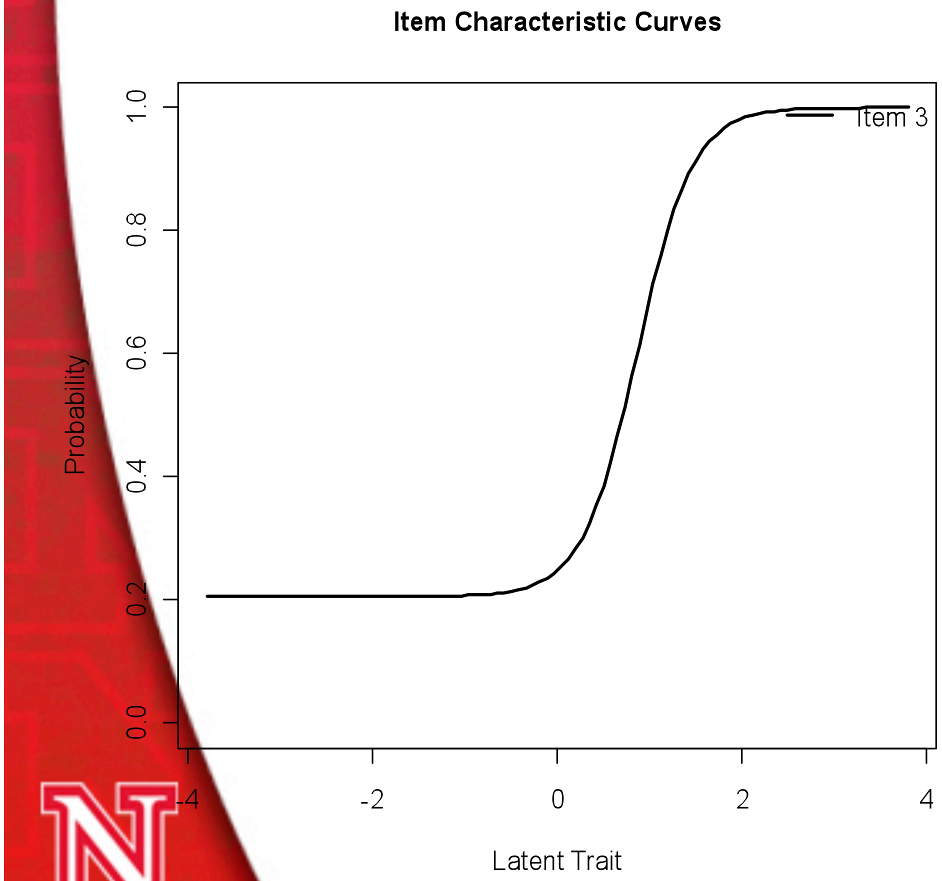
$$= c_j + (1 - c_j)\frac{\exp[Da_j(\theta_i - b_j)]}{1 + \exp[Da_j(\theta_i - b_j)]}$$

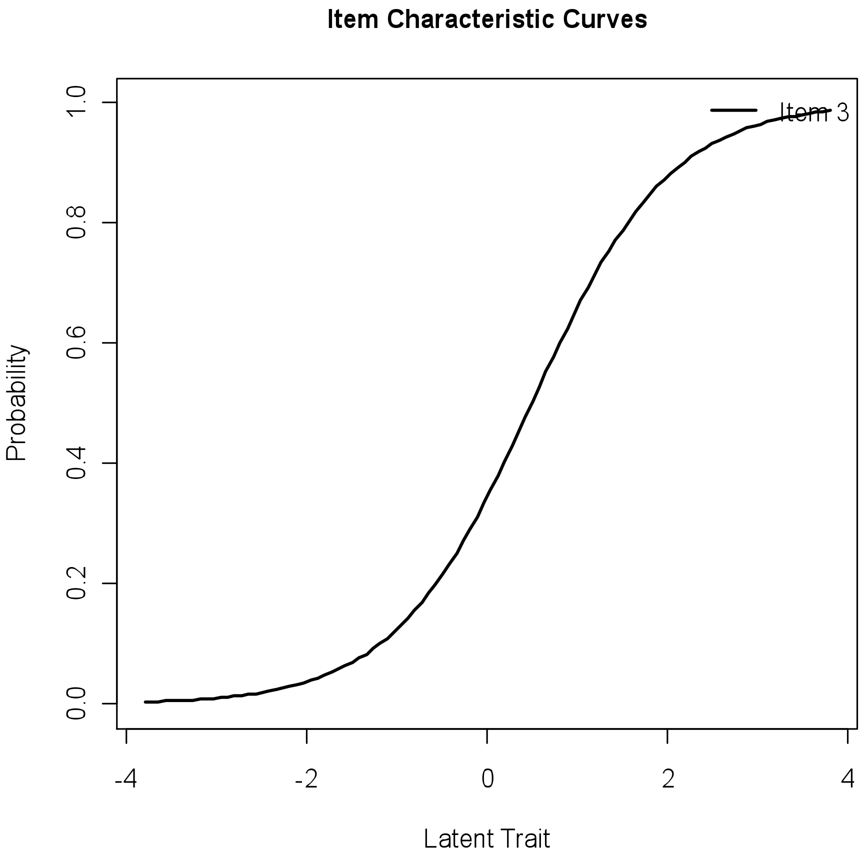
where θ_i is an examinee's ability on a single construct, a_j is the item discrimination, b_j is the item difficulty, c_j is the lower asymptote (guessing parameter), and *D* is a scaling constant.

- 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

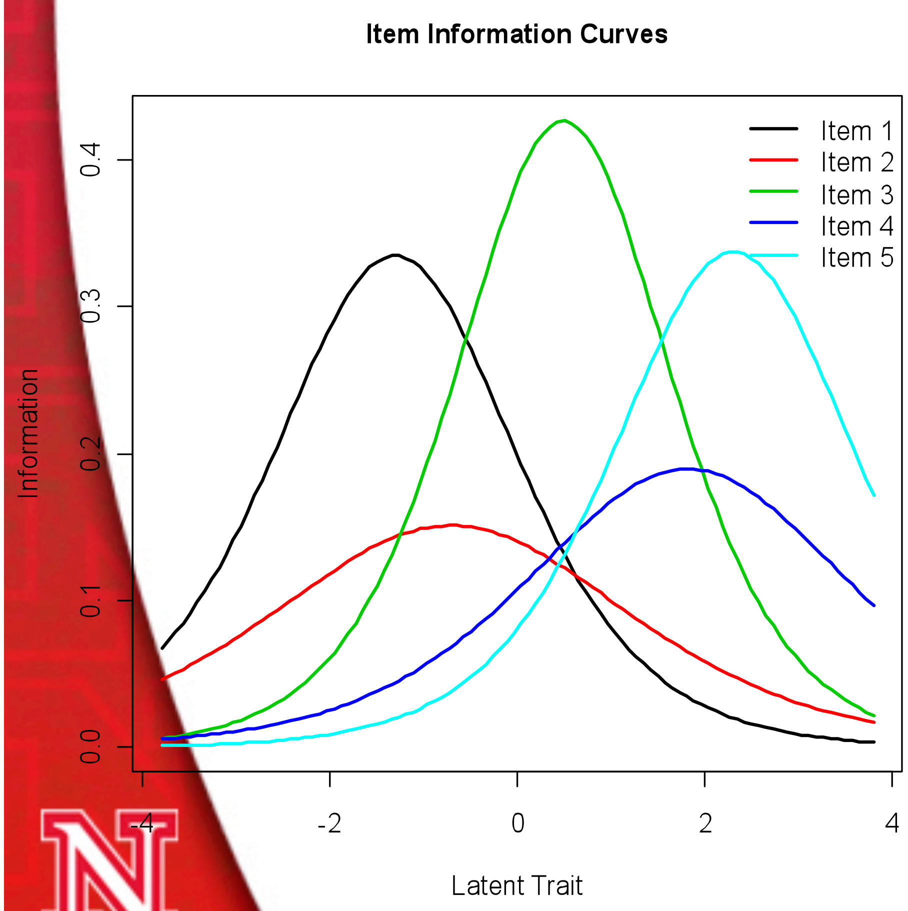
- 3PL vs. 2PL

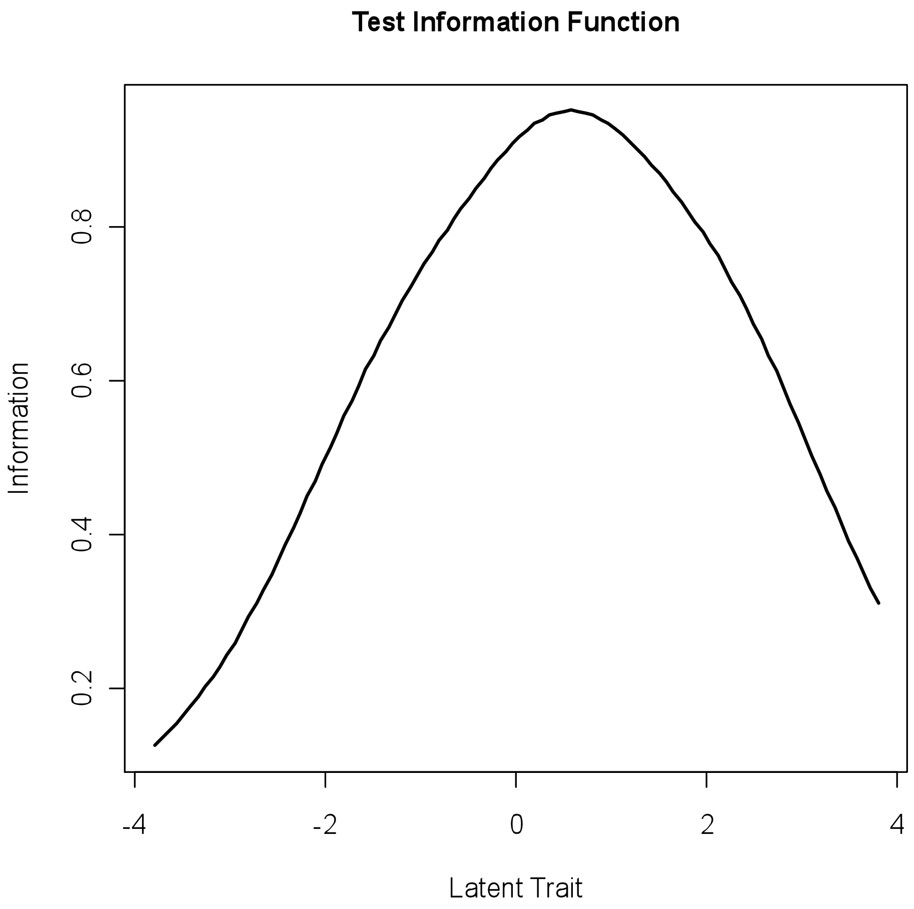




Item & Test Information Curves

- Item information curves vs. Test information curves





Polytomous IRT Models (1)

- The graded response model (GRM; Samejima 1969)

$$P_{ijk} = \widetilde{P}_{ijk} - \widetilde{P}_{ij(k+1)}$$

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

$$= \begin{cases} 1, & if \quad k = 1 \\ \dfrac{\exp[Da_j(\theta_i - b_{jk})]}{1 + \exp[Da_j(\theta_i - b_{jk})]}, & if \quad 2 \le k \le K_j \\ 0, & if \quad k > K_j \end{cases}$$

- Models
  - Constrained version: $a_i = a$ for all items
  - Unconstrained version: different a_i per item

Polytomous IRT Models (2)

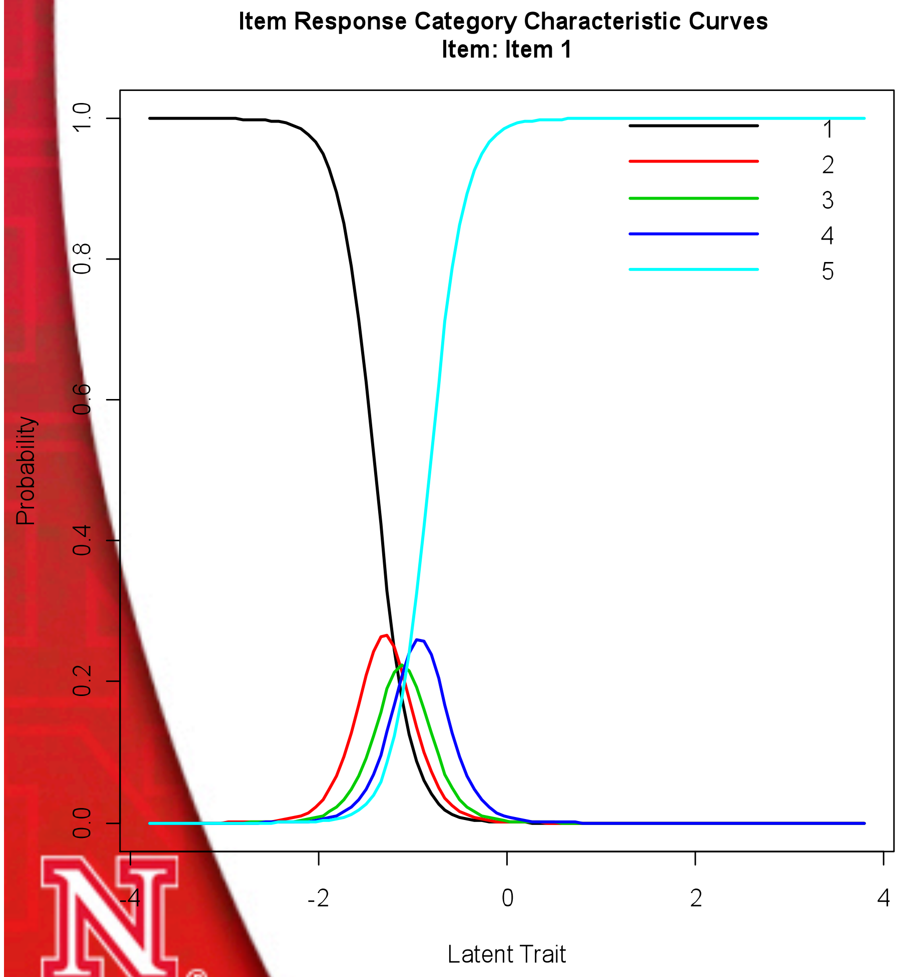
- The generalized partial credit model (GPCM; Muraki 1992)

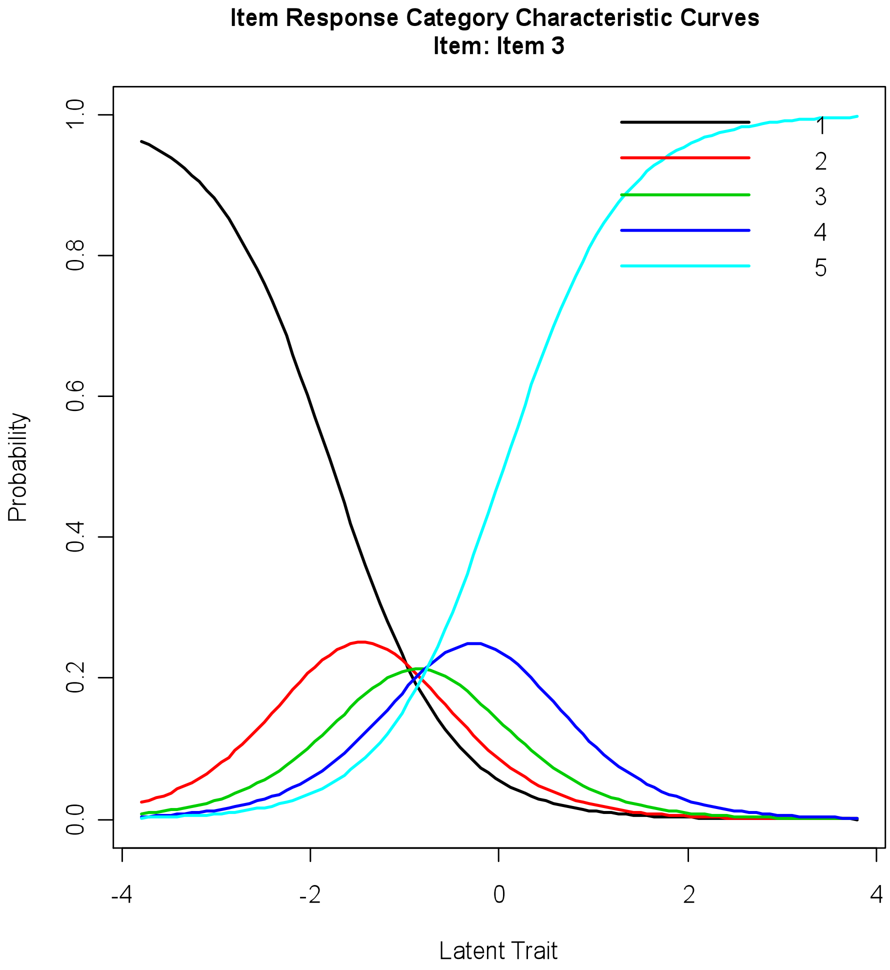
$$P_{ijk} = P(X_{ij} = v \mid \theta_i, a_j, b_{jk}) = \frac{\exp\left[\sum_{v=1}^{k} Da_j(\theta_i - b_{jv})\right]}{\sum_{h=1}^{K_j} \exp\left[\sum_{v=1}^{h} Da_j(\theta_i - b_{jv})\right]}$$

- Models
  - Constrained version: $a_i = a$ for all items
  - Unconstrained version: different a_i per item

Item Characteristic Curve (ICC)

- ICCs (Unconstrained)

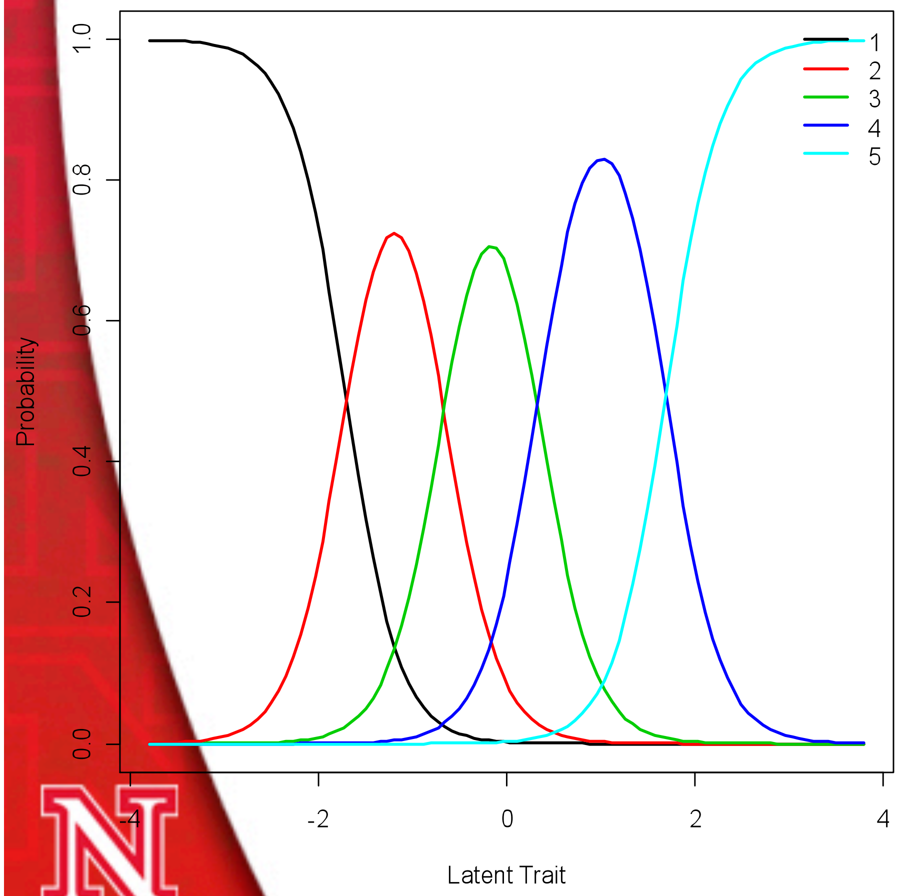

Item Response Category Characteristic Curves
Item: Item 1


Item Response Category Characteristic Curves
Item: Item 3

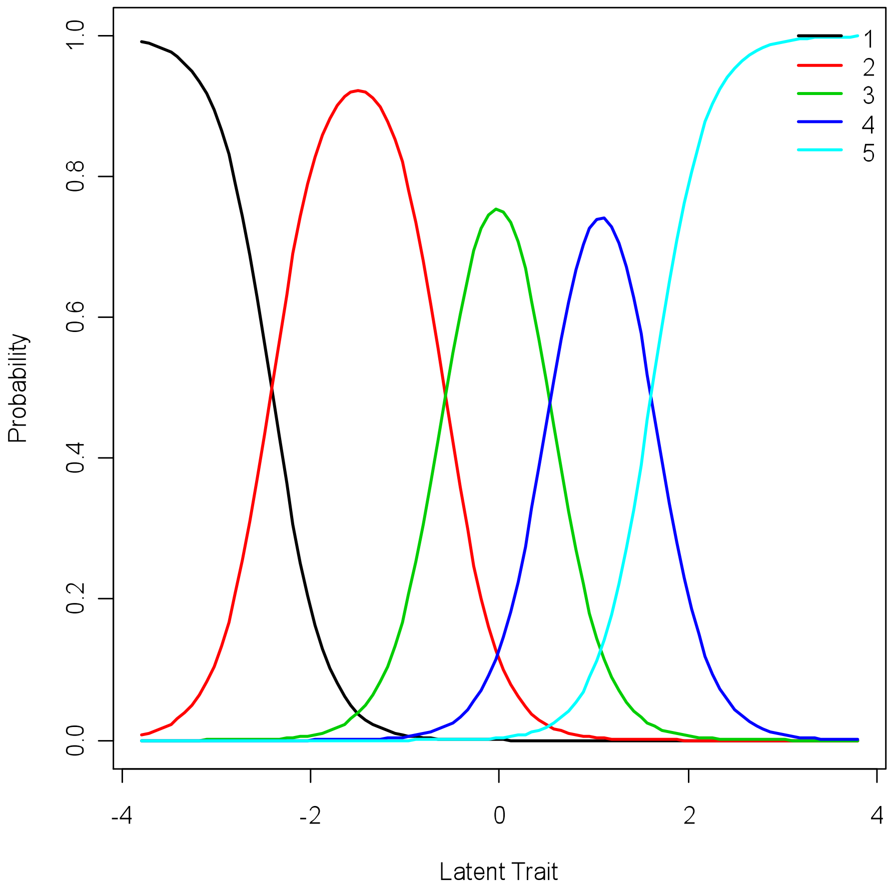
Item Characteristic Curve (ICC)

- ICCs (Constrained)

Item Response Category Characteristic Curves
Item: Item 1

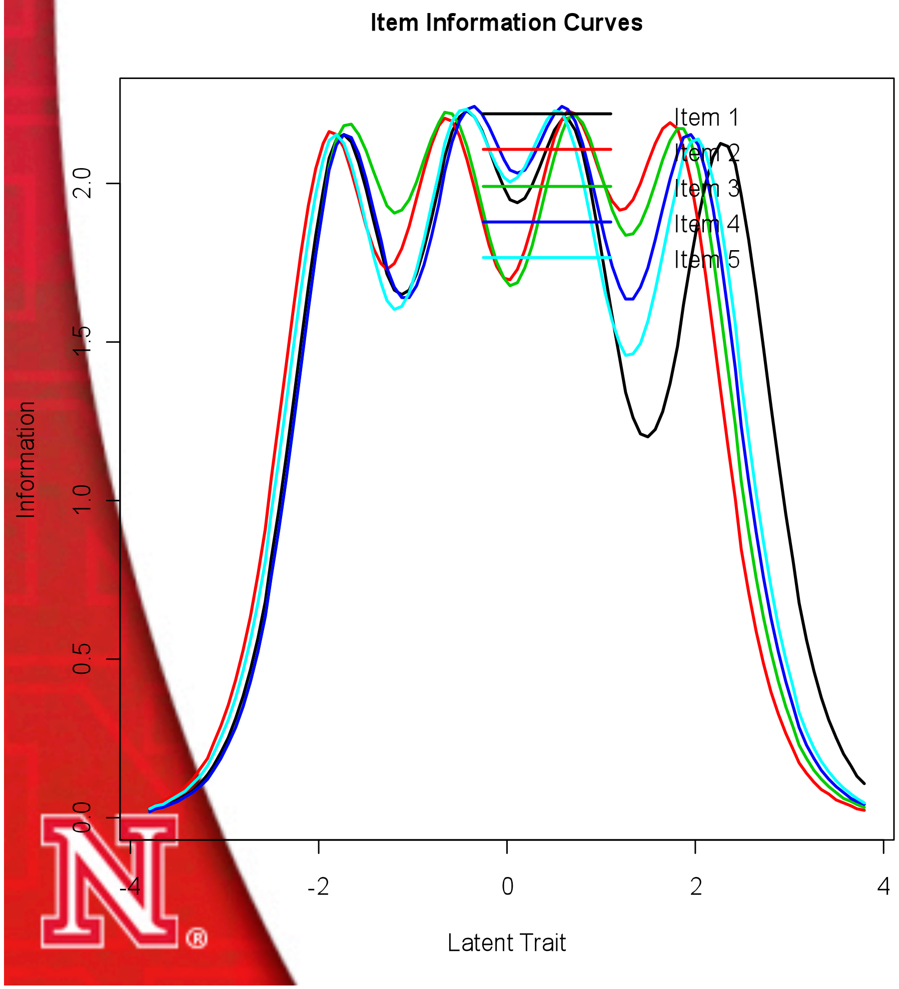


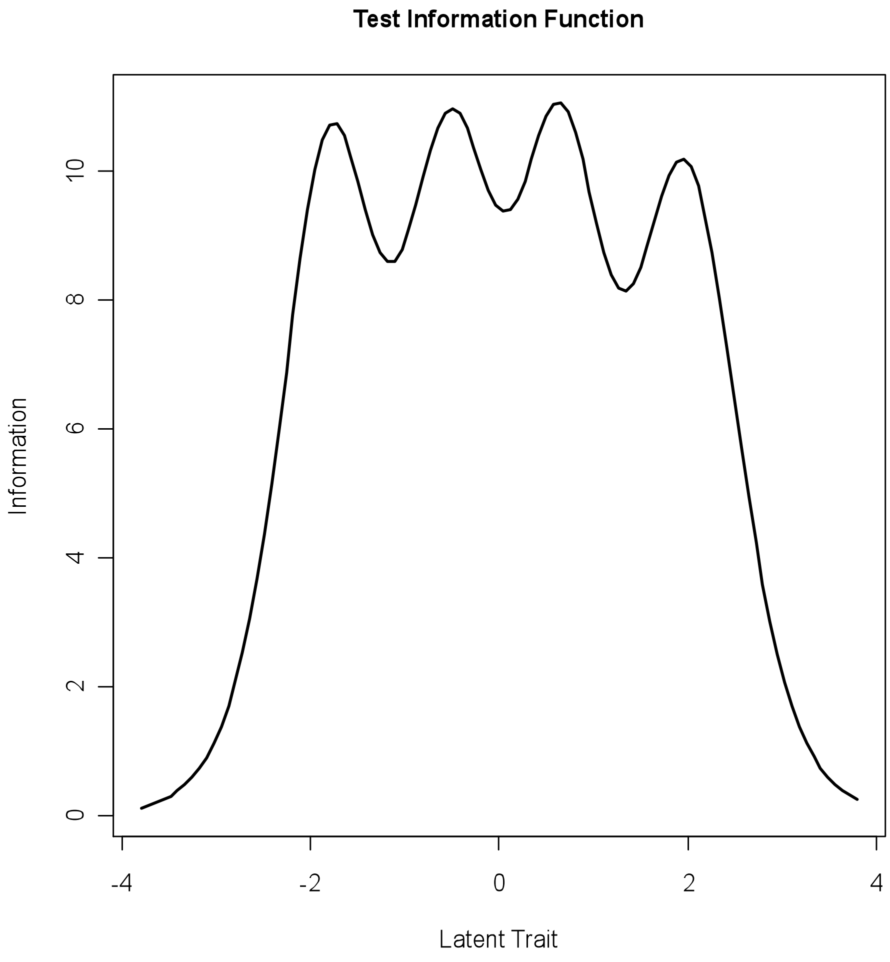
Item Response Category Characteristic Curves
Item: Item 3



Item and Test Information Curves

- Test consisting of 5 items whose number of categories in each item is 5


Item Information Curves


Test Information Function

Longitudinal Invariance (1)

- Step 1:
  - Configural invariance: test the invariance of the dimensionality
    - Pairwise association (Non-significant → 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

Subscales	Parent	Teacher
Joining (items 1 to 19)	Time 1 to Time 4	Time 1 to Time 4
Communication to others (items 20 to 24)	Time 1 to Time 4	Time 1 to Time 4

PTRS (Parent) - Joining subscale

- Configural and longitudinal metric invariance results

	Time 1	Time 2	Time 3	Time 4
Sample size (# of items)	183 (17)	159 (8)	119 (11)	87 (5)
Configural Invariance				
Scree Plot (1st eigenvalue)	One (6.840)	One (3.892)	One (3.781)	One (2.756)
Pairwise Association (items)	Items 5 & 7	Items 15, 9, 14, 11, 8, 12, & 19	Items 19, 2, 8, & 4	Items 11, 4, 8, & 14
Cronbach's α	0.8989	0.8326	0.8811	0.8989
No response on a category	Item 5	Items 5, 6, 7, &10	Items 5, 6, 12, & 13	Items 3, 5, 6, 7, 12, 13, 15, 17, 18, & 19
Metric Invariance				
IRT model (equivalent)	GRM2 (GRM1)	GRM2 (GRM1)	GRM2 (GRM1)	GRM1 (GRM2)

PTRS (Teacher) - Joining subscale

- Configural and longitudinal metric invariance results

	Time 1	Time 2	Time 3	Time 4
Sample size (# of items)	171 (12)	184 (10)	131 (11)	104 (13)
Configural Invariance				
Scree Plot (1st eigenvalue)	One (6.106)	One (5.522)	One (6.124)	One (7.943)
Pairwise Association (items)	Item 19	No item	Items 14 & 2	Items 19 & 14
Cronbach's α	0.9053	0.9053	0.9053	0.9414
No response on a category	Items 1, 3, 5, 7, 13, & 17	Items 1, 2, 3, 6, 7, 9, 11, 13, & 19	Items 1, 3, 6, 7, 13, & 19	Items 1, 2, 3, & 9
Metric Invariance				
IRT model (equivalent)	GRM2 (GRM1)	GRM2 (GRM1)	GRM2 (GRM1)	GRM2 (GRM1)

PTRS (Parent) - Comm. to others subscale

- Configural and longitudinal metric invariance results

	Time 1	Time 2	Time 3	Time 4
Sample size (# of items)	200 (5)	169 (4)	125 (5)	90 (5)
Configural Invariance				
Scree Plot (1st eigenvalue)	One (3.884)	One (3.061)	One (3.414)	One (3.563)
Pairwise Association (items)	No item	No item	No item	No item
Cronbach's α	0.9274	0.8939	0.8811	0.8989
No response on a category	No item	Item 23	No item	No item
Metric Invariance				
IRT model (equivalent)	GRM1	GRM1 (GPCM2)	GRM1	GRM1

PTRS (Teacher)
- Comm. to others subscale

- Configural and longitudinal metric invariance results

	Time 1	Time 2	Time 3	Time 4
Sample size (# of items)	197 (5)	192 (3)	137 (4)	107 (4)
Configural Invariance				
Scree Plot (1st eigenvalue)	One (3.465)	One (2.071)	One (2.973)	One (2.644)
Pairwise Association (items)	No item	No item	No item	No item
Cronbach's α	0.8842	0.7622	0.8278	0.8217
No response on a category	No item	Items 20 & 23	Item 21	Item 20
Metric Invariance				
IRT model (equivalent)	GRM1 (GRM 2)	GRM2 (GPCM2)	GRM1 (GRM2)	GRM 2

Results on Longitudinal Metric Invariance

- List of longitudinal invariant items and selected models for subscales

	Parent		Teacher	
	Joining	Communication to others	Joining	Communication to others
Longitudinal invariant items	**2 out of 19 items** (10.5%)	**4 out of 5 items** (80%)	**7 out of 19 items** (36.8%)	**1 out of 5 items** (20%)
IRT model	GRM without constraints	GRM with constraints	GRM without constraints	GRM without constraints

- Extreme measurements and common items used in linking process

	Parent - Joining	Parent - Comm. to others
Common items	**1 and 16**	**20, 21, 22, and 24**
Percent of total information provided by longitudinal invariant items	Time 1 (13.87%)	Time 1 (78.86%)
	Time 2 (30.21%)	Time 2 (100%)
	Time 3 (21.79%)	Time 3 (79.72%)
	Time 4 (39.93%)	Time 4 (82.38%)

Results on Linking

- Linking constants for parent joining subscale

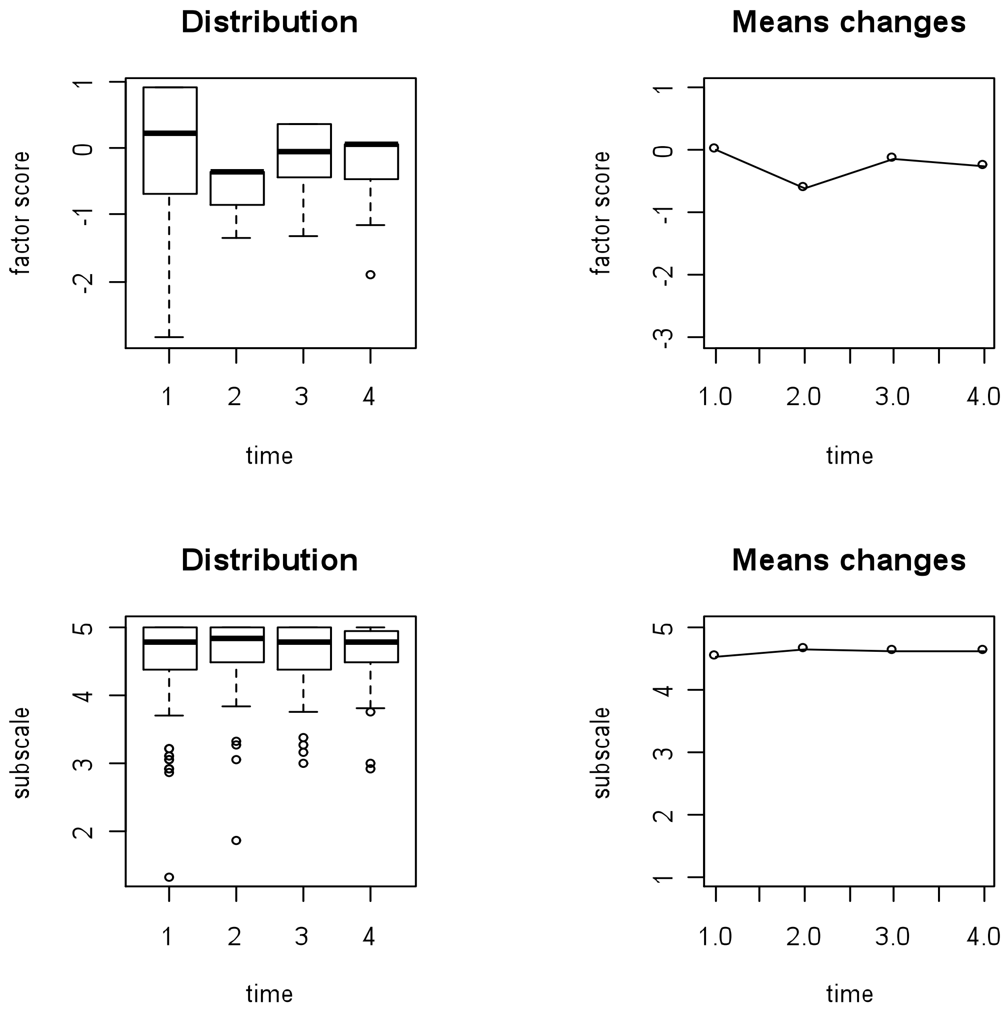
	Time 2 to Time 1		Time 3 to Time 1		Time 4 to Time 1	
	A	B	A	B	A	B
Haebara	0.274	-0.567	0.543	-0.018	0.565	-0.115
Stocking-Lord	0.381	-0.258	0.532	-0.054	0.657	0.032

- Linking constants for parent communication to others subscale

	Time 2 to Time 1		Time 3 to Time 1		Time 4 to Time 1	
	A	B	A	B	A	B
Haebara	0.705	-0.009	0.661	-0.088	0.454	-0.252
Stocking-Lord	0.810	0.181	0.755	0.073	0.585	-0.042

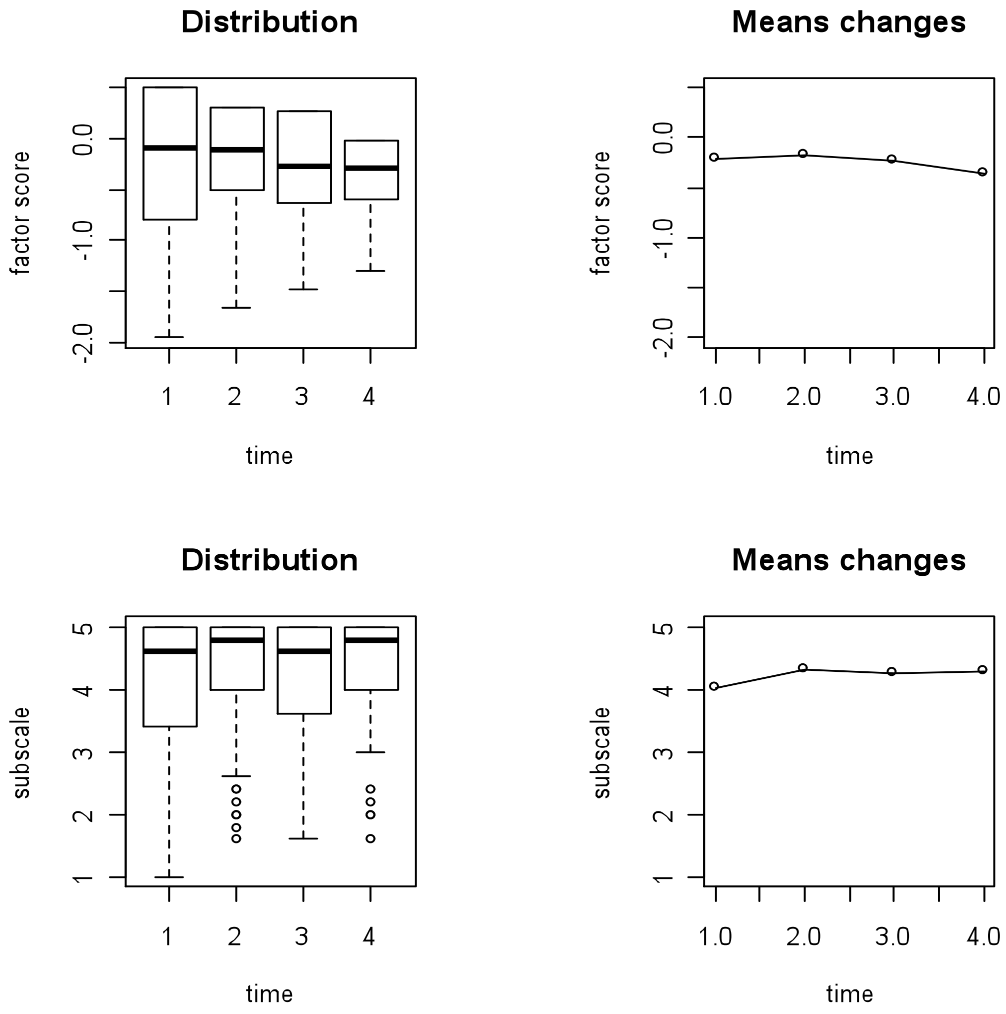
Change over time on joining subscale

- 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

		AIC	BIC	χ^2	p-value
Factor	Constant	606.88	618.11		
	Log	592.88	607.85	16.00	0.00
	Treatment	591.67	610.39	3.20	0.07
Subscale	Constant	469.15	480.38		
	Linear	470.18	488.89	2.97	0.23
	Treatment	466.25	481.22	4.90	0.03

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

1. Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In FM Lord, MR Novick (eds.), *Statistical Theories of Mental Test Scores,* 397-479. Addison-Wesley, Reading, MA.
2. de Frias, C.M. and Dixon, R.A. (2005). Confirmatory factor structure and measurement invariance of the memory compensation questionnaire. *Psychological Assessment, 17*, 168-178.
3. Horn, J.L. and McArdle, J.J. (1992). A practical and theoretical to measurement invariance in aging research. *Experimental Againg Research, 18*, 117-144.
4. Maitland, S.B., Dixon, R.A., Hultsch, D.F. and Hertzog, C. (2001). Well-being as a moving target: Measurement equivalence of the Bradburn Affect Balance Sheet. *Journal of Gerrontology, 56B*, 69-77.
5. Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 525-542.
6. Molt, R.W., Dishman, R.K., Birnbaum, A.S., and Lytle, L.E. (2005). Longitudinal invariance of the center for epidemiological studies-depression scale among girls and boys in middle school. *Educational and Psychological Measurement, 65*, 90-108.
7. Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16(2)*, 159-176.
8. Obradovic, J., Pardini, D.A., Long, J.D., and Loeber, R. (2007). Measuring Interpersonal Callousness in Boys From Childhood to Adolescene: An Examination of Longitudinal Invariance and Temporal Stability. *Journal of Clinical Child and Adolescent Psychology, 36(3)*, 276-292.
9. Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17(5)*, 1-25.
10. Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement*, *34,* 100-114.
11. Vickers, H.S. and Minke, K.M. (1995). Exploring parent teacher relationships: Joining and communication to others. *School Psychology Quarterly, 10(2)*, 133-150.
12. Week, J.P. (2010) plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software, 35(12)*, 1-3.

Thank you!

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

University of
Nebraska
Lincoln