Metric Misspecification due to Test Multidimensionality and Consequences for the Measurement of Growth

Xiangyi Liao[1] Daniel M. Bolt[1] Jee-Seon Kim[1]

[1]University of Wisconsin, Madison

Ideas In Testing Research Seminar
November 4, 2022

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Critiques of the Interval Property of IRT Metric

- Educational research outcomes frequently rely on an assumption that measurement metrics have interval-level properties.
- Education measurement scales, including the latent scales derived from item response theory (IRT) models, may lack interval scale properties that permit comparisons of score gains (Ballou, 2009; Betebenner, 2011; Michell, 2009).
- While most investigators know enough to be suspicious of interval-level claims, and in some cases even question their findings in light of such suspicions, what is absent is an understanding of the measurement conditions that create metric distortions.

ECLS-K Framework & Measurement Modeling

- ECLS-K (Early Childhood Longitudinal Study) Reading Assessment
  - possible dimensionality issue in the test items
    e.g. items on sub-domains including basic skills, initial understanding, developing interpretaion, and critical stance
  - dimensionality is related to different item types
    e.g. "name letter" (easier) versus "decoding" (more difficult) items
  - dimensions are highly correlated
  - Unidimensional Item Response Theory (UIRT) model is used to scale the test scores
- We seek to simulate multidimensionality of the form on ECLS-K and examine metric distortion when 2PL is applied.

Fitting UIRT to Multidimensional Data: Model Fit

- Two-dimensional response data
  - highly correlated dimensions

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \sim N\left(\begin{pmatrix}-1\\ 1\end{pmatrix}, \begin{pmatrix}1 & 0.8\\ 0.8 & 1\end{pmatrix}\right)$$

  - Between-item dimensionality
  - easy items measuring θ_1, difficult items measuring θ_2
- Model fit

	AIC	BIC	logLik
UIRT	197711.2	198232.6	-98775.62
MIRT	195555.9	196083.8	-97696.95

- Item-fit statistics can hardly detect any misfit when fitting UIRT model to multidimensional data with highly correlated dimensons.

Fitting UIRT to Multidimensional Data

A long-standing conjecture: the fitted UIRT to multidimensional data represents a linear composite of the dimensions present in a test.

$$\theta_\alpha = w_1\theta_1 + w_2\theta_2$$

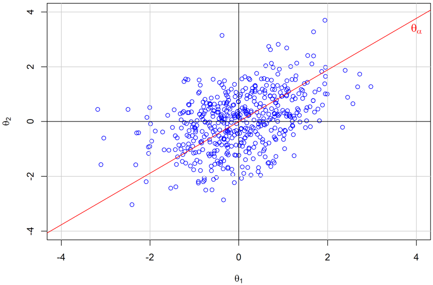

Figure 1: Illustration of a latent bivariate distribution for (θ_1, θ_2) with a corresponding linear composite direction denoted by θ_α, Strachan et al. (2022)

UIRT Approximation: Two-dimensional Example

- Two-dimensions where dimensionality is related to item types

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \sim N\left(\begin{pmatrix}-1\\ 1\end{pmatrix}, \begin{pmatrix}1 & 0.8\\ 0.8 & 1\end{pmatrix}\right)$$

  - θ_1 on easy items: $a \sim N(1.3, 0.2)$,$b \sim N(-1, 1)$
  - θ_2 on difficult and discriminative items: $a \sim N(2, 0.2)$,$b \sim N(1, 1)$
- Calibrate the response data with UIRT model
- Estimate w_1 and w_2 by ability groups from separate latent regressions

$$\hat{\theta} = \hat{w}_1\theta_1 + \hat{w}_2\theta_2$$

UIRT Approximation: Two-dimensional Example, Cont'd

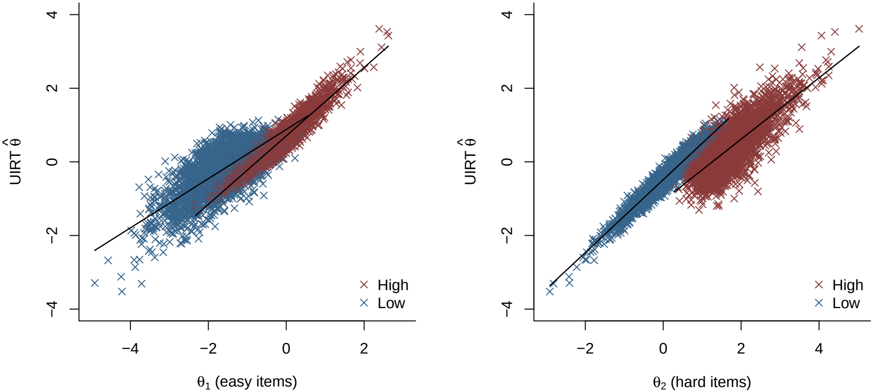

Figure 2: Illustration of the UIRT Approximation by Dimension, Two Group

UIRT Approximation: Two-dimensional Example, Cont'd

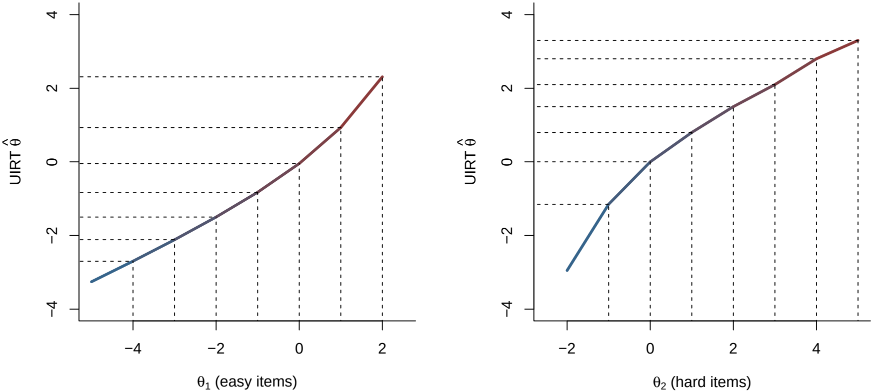

Figure 3: Illustration of the UIRT Approximation by Dimension

UIRT Approximation: Two-dimensional Example, Cont'd

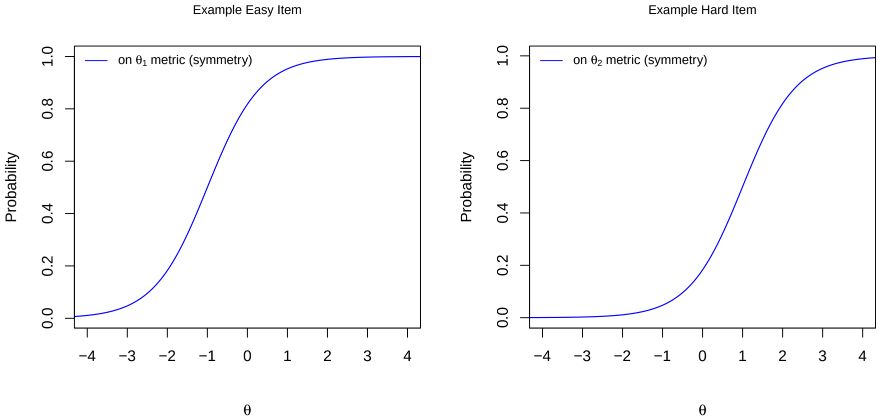

Figure 4: Illustration of the UIRT Metric Distortion in ICCs

UIRT Approximation: Two-dimensional Example, Cont'd

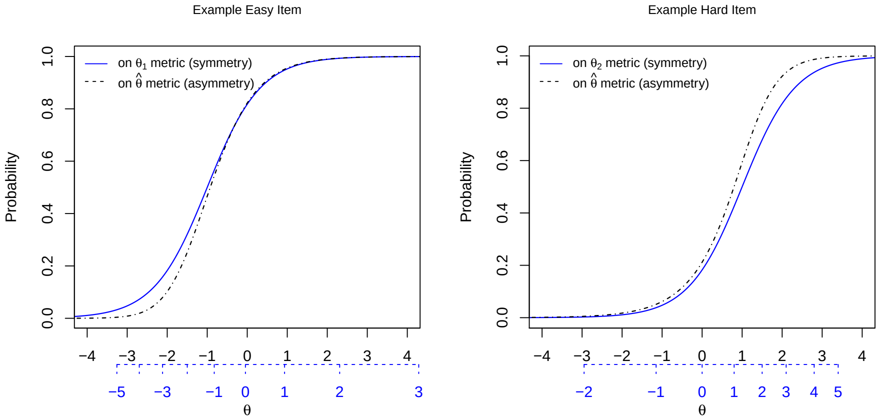

Figure 5: Illustration of the UIRT Metric Distortion in ICCs

Consequences of UIRT Approximation: Matthew Effect

- Students who start lower on the metric may tend to be credited with lesser gains than students that start higher even if they grow equivalent amounts.

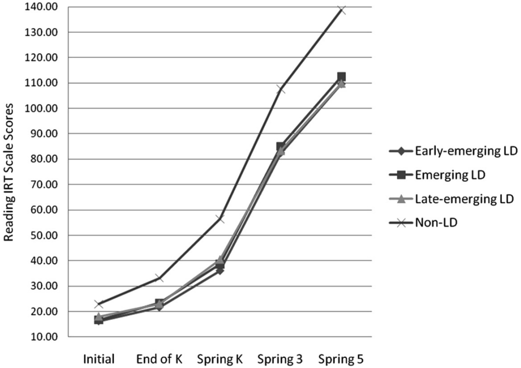

Figure 6: Group Differences in Reading Growth and Achievement over the First 6 Years of School, ECLS-K data, from Judge & Bell (2010)

Consequences of UIRT Approximation: Vertical Linking

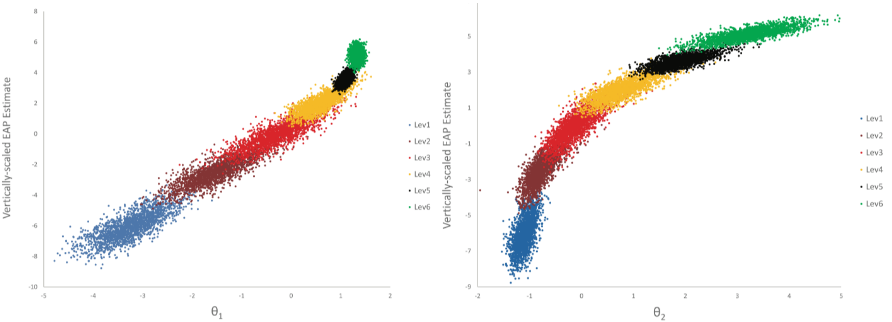

Figure 7: Relationships between Vertically Scaled EAP Estimates and θs, from Carlson (2017)

Conclusion

- UIRT θ as a curvilinear approximation when dimensionality is related to item difficulty
- Interpretation of the UIRT θ
  "The IRT scale scores may be used as longitudinal measures of overall growth. However, gains made at different points on the scale have qualitatively different interpretations. [...] Comparison of gain in scale score points is most meaningful for groups that started with similar initial status." (Pollack et al., 2005)
- Selecting anchor items in vertical linking

References

Ballou, D. (2009). Test scaling and value-added measurement. *Education finance and Policy*, 4(4):351–383.

Betebenner, D. W. (2011). A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories. *National Center for the Improvement of Educational Assessment*.

Carlson, J. E. (2017). Unidimensional vertical scaling in multidimensional space. *ETS Research Report Series*, 2017(1):1–28.

Judge, S. and Bell, S. M. (2010). Reading achievement trajectories for students with learning disabilities during the elementary school years. *Reading & Writing Quarterly*, 27(1-2):153–178.

Michell, J. (2009). The psychometricians' fallacy: too clever by half? *British Journal of Mathematical and Statistical Psychology*, 62(1):41–55.

Pollack, J. M., Rock, D. A., Weiss, M. J., Atkins-Burnett, S., Tourangeau, K., West, J., and Hausken, E. G. (2005). Early childhood longitudinal study, kindergarten class of 1998-99 (ecls-k): Psychometric report for the third grade. nces 2005-062. *National Center for Education Statistics*.

Samejima, F. (2000). Logistic positive exponent family of models: Virtue of asymmetric item characteristic curves. *Psychometrika*, 65(3):319–335.

Strachan, T., Cho, U. H., Ackerman, T., Chen, S.-H., de la Torre, J., and Ip, E. H. (2022). Evaluation of the linear composite conjecture for unidimensional irt scale for multidimensional responses. *Applied Psychological Measurement*, 46(5):347–360.

Thank you!

Xiangyi Liao	xliao36@wisc.edu
Daniel Bolt	dmbolt@wisc.edu
Jee-Seon Kim	jeeseonkim@wisc.edu

S1. ICC Asymmetry

- Samejema's (2000) logistic positive exponent (LPE) model

$$P_{ij}(X_{ij} = 1|\theta_i; a_j, b_j, \xi_j) = \left(\frac{\exp\left(a_j(\theta_i - b_j)\right)}{1 + \exp\left(a_j(\theta_i - b_j)\right)}\right)^{\xi_j}$$

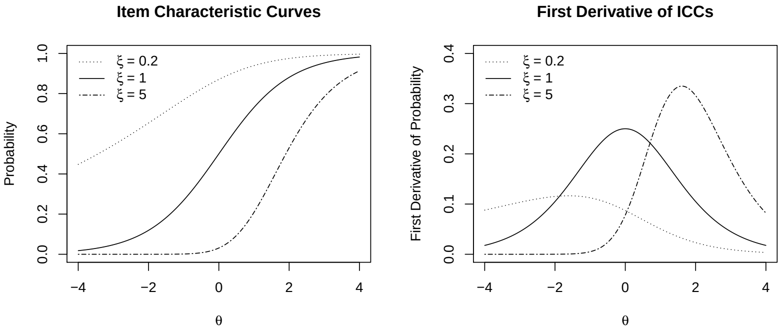

Figure 8: Example Item Characteristic Curves (ICCs) and their First Derivatives of LPE Items ($a = 1, b = 0$ for all items).