

Differential Item Functioning by Multiple Variables using Moderated Nonlinear Factor Analysis

Sanford R. Student & Ethan M. McCormick

About the Authors

Sanford (Sandy) Student is an Assistant Professor of Educational Statistics and Research Methods, and resident faculty of the Data Science Institute, at the University of Delaware. He teaches courses in survey design, educational measurement, Item Response Theory, and structural equation modeling. His research focuses on the relationship between psychometric issues, particularly in the measurement of growth using vertical scales, and broader inferences about student learning made on the basis of test scores. This includes research on the psychometric methods and data collection designs used in the development of vertical scales; research on the use of vertically scaled test scores in the educational evaluation literature, such as the estimation of the magnitude of a “year of learning” for different subjects and grades; and using vertically scaled test scores in multilevel and structural equation approaches to modeling academic growth over time. He is currently the Principal Investigator of an Institute of Education Sciences-funded grant investigating the use of Moderated Nonlinear Factor Analysis to strengthen cross-grade comparisons on a vertical scale, and his work has appeared in journals such as *Educational Researcher*, *Educational Measurement: Issues and Practice*, and the *Journal of Research on Educational Effectiveness*. He is an active member of NCME, having joined the organization early in graduate school.



About the Authors

Ethan McCormick is an assistant professor of Educational Statistics and Research Methods, and resident faculty of the Data Science Institute, at the University of Delaware. He was previously at Leiden University in the Netherlands. He received his Ph.D. from the University of North Carolina at Chapel Hill. His research and teaching interests include specializing in longitudinal, time series, and psychometric models, focusing on structural equation modeling, multilevel modeling, Bayesian hierarchical time series modeling, and nonlinear modeling approaches for understanding population heterogeneity in measurement and change over time. His work has appeared in a variety of leading methodological journals, including *Psychometrika* and *Psychological Methods*, as well as a line of translational research applying advanced psychometric models in neuroscience, with work appearing in journals such as *Journal of Neuroscience* and *Network Neuroscience*. In addition to his research, he has conducted a variety of in-person and remote workshops to leading universities in the US, UK, EU, and Australia, on topics including: multilevel modeling, structural equation modeling, longitudinal modeling, and nonlinear approaches to measurement.



Learning Objectives

1 Articulate the difference between uniform and non-uniform DIF in the slope-intercept form of the 2PL IRT model.

2 Differentiate DIF from impact, and describe the implications of both for parameters of traditional IRT models.

3 Describe how moderated nonlinear factor analysis can be applied to estimate both DIF and impact in the slope-intercept 2PL.

4 Apply regularized moderated nonlinear factor analysis to simultaneously estimate DIF and impact for multiple covariates of mixed types (i.e. categorical and continuous) using the R package regDIF.

5 Use the results of regularized MNLFA estimation to inform next steps in DIF analysis.

Module Sections

- Introduction
- Section 1: IRT, DIF, and a More General Model
- Section 2: An Overview of MNLFA for DIF and Impact Assessment
- Section 3: MNLFA Estimation and Interpretation
- Section 4: MNLFA Applied Example
- Section 5: MNLFA Code Walkthrough
- Accompanying online guided activity, dataset, and list of additional readings/resources

IRT, DIF, and a More General Model

1

Section Learning Objectives

1 IRT, DIF, and a more general model

Recognize the isomorphism between the 2PL IRT model and a slope-intercept factor analysis model

Differentiate and define the meaning of DIF versus impact

Contrast the strengths and limitations of multigroup and MIMIC approaches to DIF and impact

Recognize how MNLFA generalizes the slope-intercept model to estimate DIF and impact

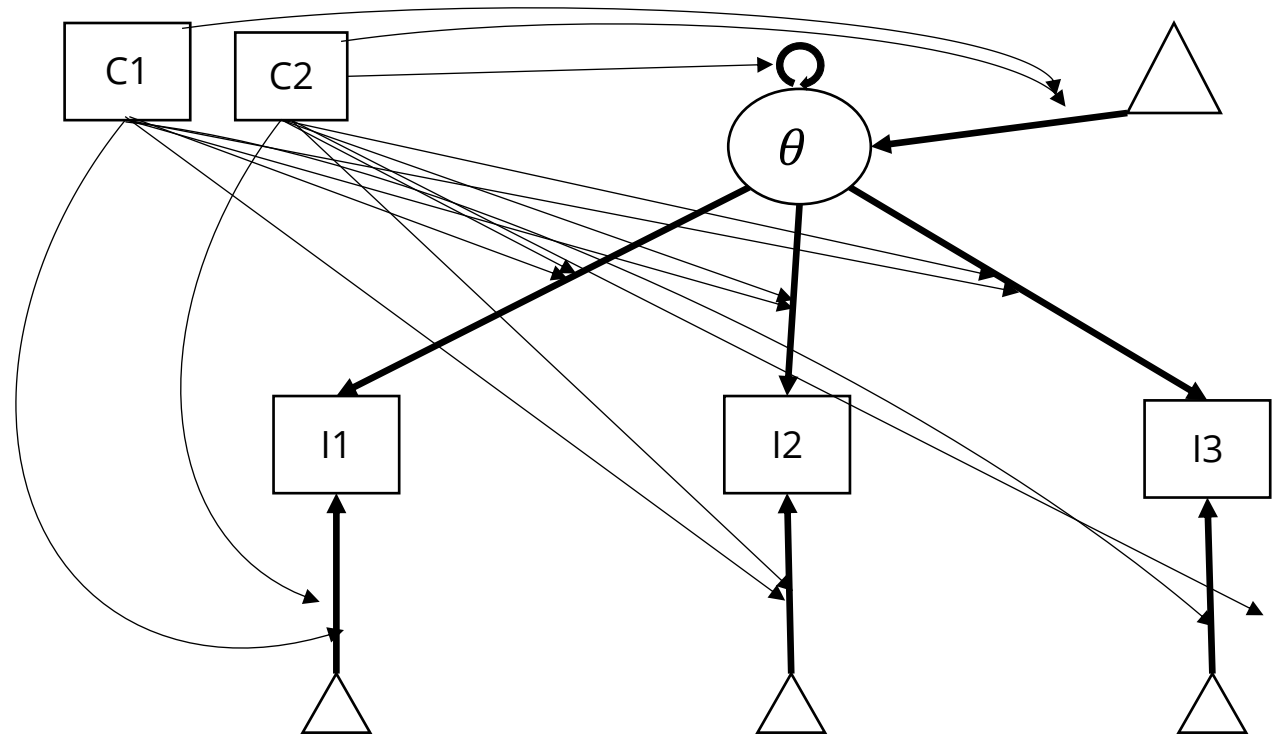
A Brief Overview

- This ITEMS module is about differential item functioning (DIF) analysis
- The variables defining subgroups are often called “background variables”
- For example, race/ethnicity, gender, English language proficiency, socioeconomic status, etc.
- Typically done two groups at a time

Most methods cannot accommodate multiple grouping variables or continuous variables

A Brief Overview

- Moderated nonlinear factor analysis (Bauer & Hussong, 2009; Bauer, 2017) is a generalized measurement model with connections to Item Response Theory and structural equation modeling that can be used to evaluate DIF much more flexibly



The Two-Parameter Logistic Model

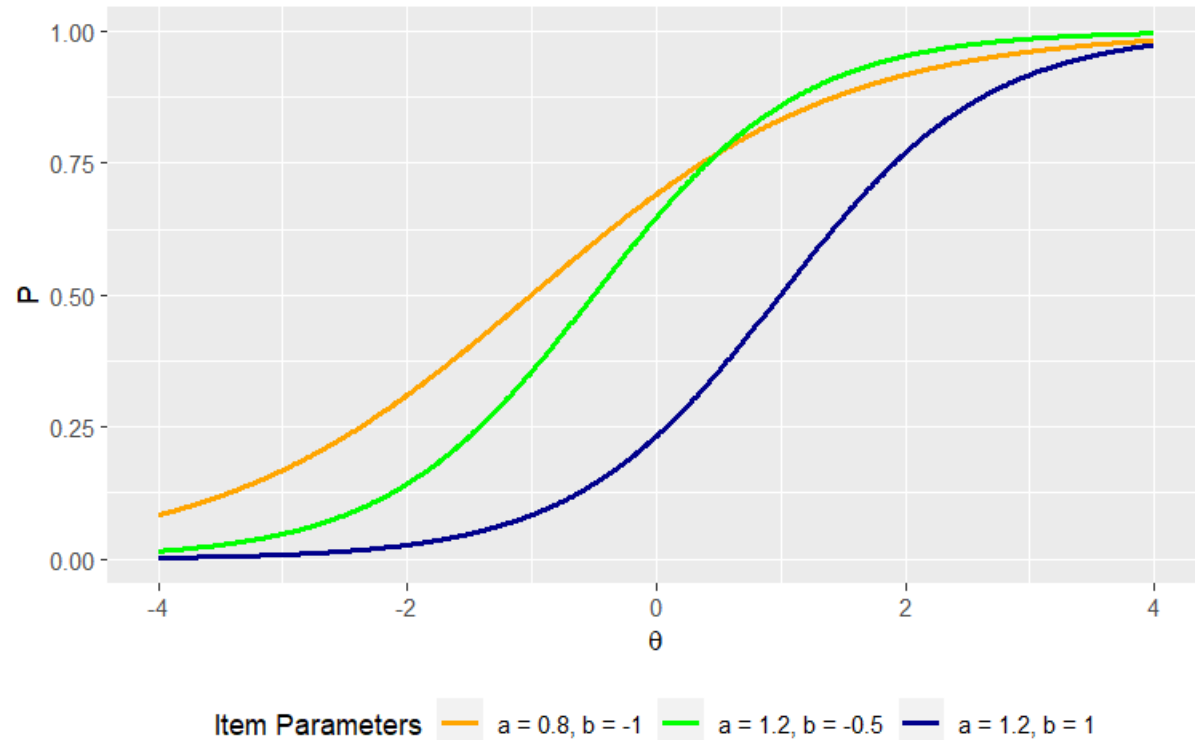
$$P(X_{ij} = 1 | \theta_i, b_j, a_j) = \frac{\exp(a_j(\theta_i - b_j))}{1 + \exp(a_j(\theta_i - b_j))}$$

- The 2PL models the probability of a correct response to an item as a function of:
 - The examinee's ability θ_i
 - The item's difficulty b_j
 - The item's discrimination a_j
 - Typically $\theta \sim N(\mu, \sigma^2)$, specifically $\theta \sim N(0,1)$

The Two-Parameter Logistic Model

Three ICCs

Under the Two-Parameter Logistic Model



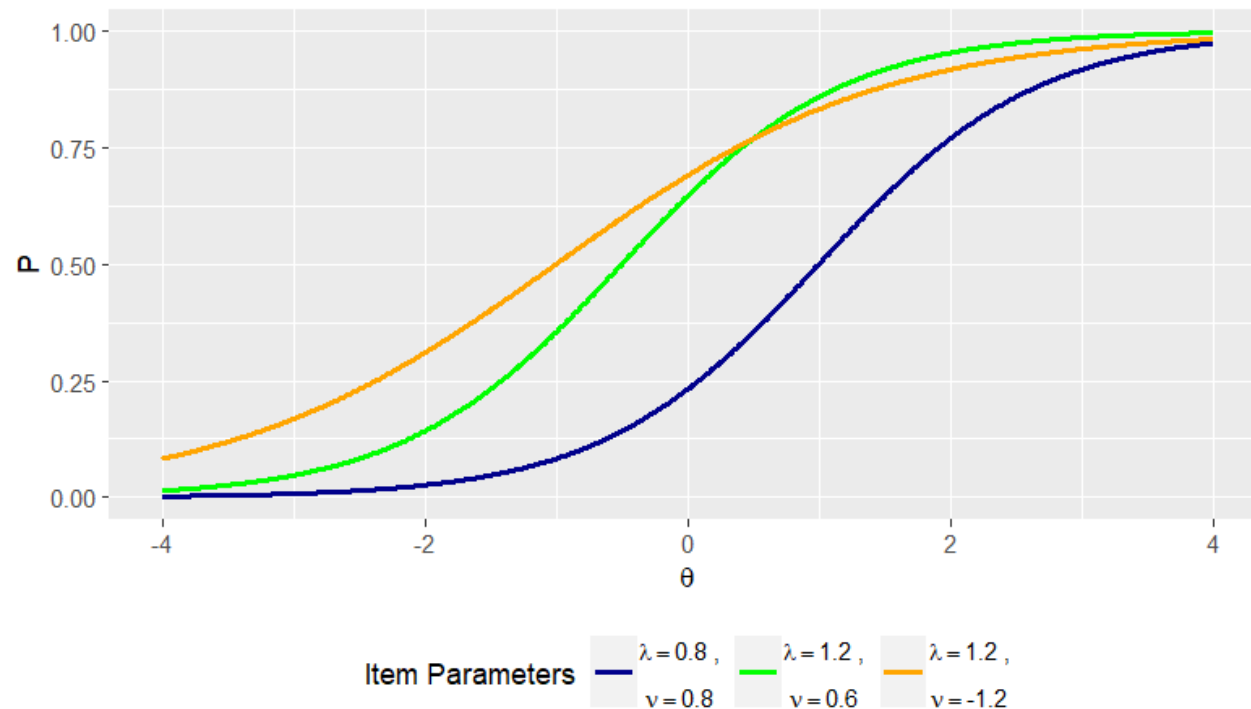
- The probability is produced by the *item response function* (IRF)
- The 2PL is for dichotomous items

The Two-Parameter Logistic Model

$$P(X_{ij} = 1) = \frac{\exp(\lambda_j \theta_i + \nu_j)}{1 + \exp(\lambda_j \theta_i + \nu_j)}$$

Three ICCs

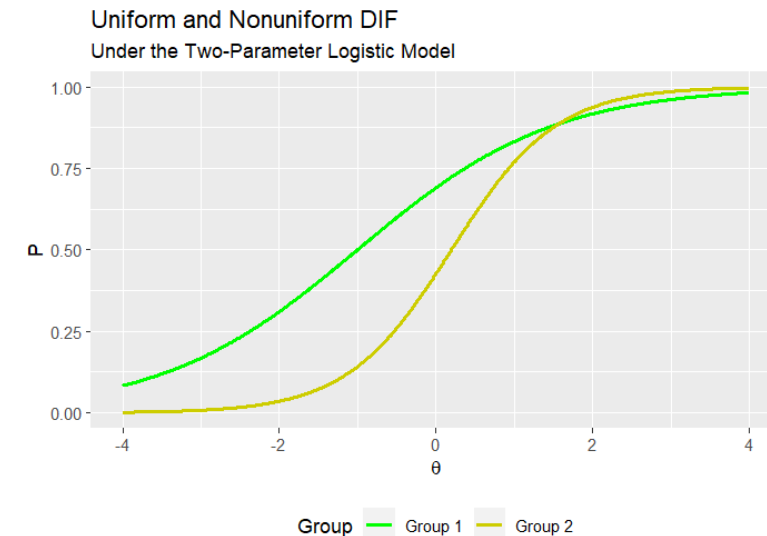
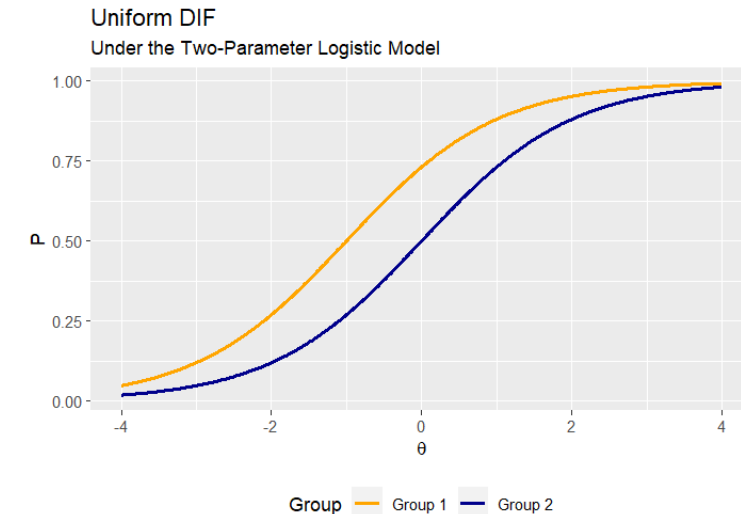
Under the Two-Parameter Logistic Model Using Slope-Intercept Parameterization



- The 2PL can be reparameterized as a linear model in the logit metric – the “slope-intercept” form – where:
 - $\lambda_j = a_j$
 - $\nu_j = -b_j a_j$
 - No change in the probabilities
 - With $\theta \sim N(0,1)$
 - See Kamata and Bauer (2008)

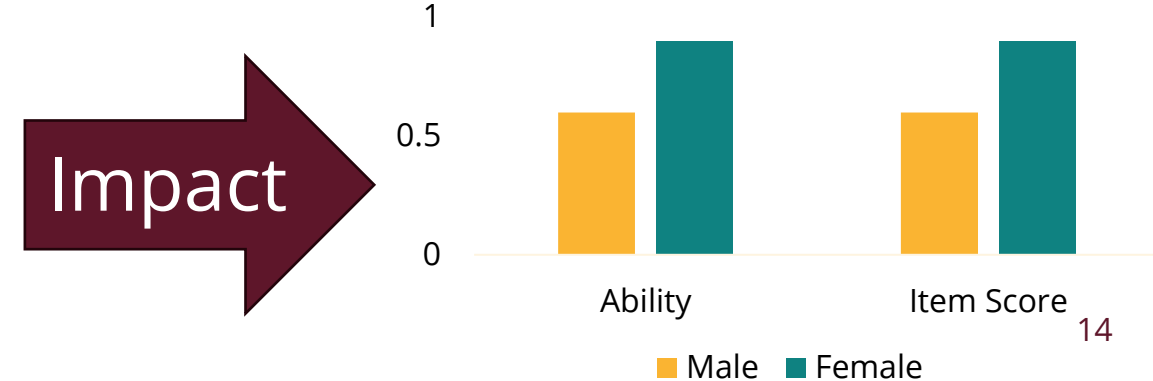
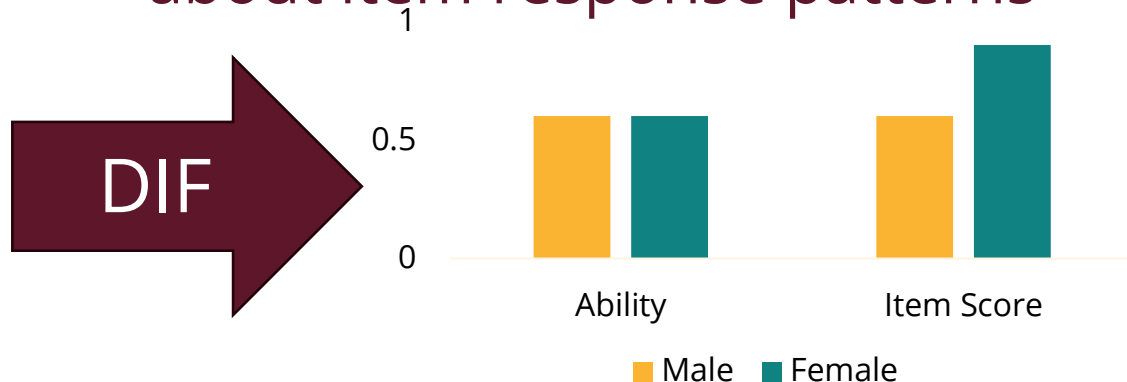
IRT Differential Item Functioning

- Lord, 1980, p. 212: “If... an item has a different item response function for one group than for another, it is clear that the item is biased.”
 - This is the IRT framing of differential item functioning (DIF).
 - See Thissen (2024) for more.



DIF vs. Impact

- DIF: an item's IRF is conditional on background variable(s)
 - For educational tests, typically interpreted as item-level *bias*
 - A phenomenon we aim to identify and *mitigate*
- Impact: θ distribution conditional on background variable(s)
 - For educational tests, typically interpreted as group-level *differences*
 - A phenomenon we aim to identify in broader contexts (e.g., policy) but generally do not aim to mitigate during test development
 - Needs to be accounted for in DIF analysis
 - After accounting for impact, group membership should tell us nothing about item response patterns



DIF vs. Impact

- DIF and impact are typically assessed two groups at a time, but this is a matter of statistical limitations of DIF methods
- One might want to evaluate DIF across more than one variable at once for several reasons, such as:
 - Intersectionality (Russell, 2024; Xu & Soland, 2024)
 - Correlated background variables and disentangling sources of DIF (e.g., race and socioeconomic status)
 - Correct model specification!

DIF and Measurement Invariance

- DIF: an item's IRF is conditional on background variable(s)
- DIF can be framed in terms of measurement invariance from structural equation modeling literature (Meredith, 1993):

$$F(\mathbf{x}|\mathbf{w}, \mathbf{v}) = F(\mathbf{x}|\mathbf{w})$$

- \mathbf{x} is an observed item response vector, \mathbf{w} is a (vector of) latent variable(s), \mathbf{v} is a (vector of) background variable(s)
- This equation states that conditional on w , item responses are unrelated to group membership.
 - \mathbf{w} in Meredith's presentation is equivalent to a vector of θ s in IRT.
 - This is another way of describing a lack of DIF.
 - If this condition does not hold, we have DIF.

Two Related Models

- IRT approach: multiple group analysis (e.g. Thissen et al., 1988)
- SEM approach: multiple indicators, multiple causes (with interaction; e.g. Woods and Grimm, 2011)

Existing Methods for IRT/SEM DIF Analysis

Multigroup IRT-based methods

- Based on IRT parameterization
- Allows for simultaneous modeling of DIF and mean/variance impact
- Categorical background variables only
- More than one grouping variable? Must be combined to create mutually exclusive groups

SEM-based MIMIC-with-interaction models

- Based on slope-intercept parameterization
- Allows for simultaneous modeling of DIF and mean impact
- Categorical, continuous, nominal, censored, etc. background variables; groupings can overlap

See Bauer (2023) for a deeper comparison of these and other models.

MNLFA: Draws on the Strengths of Both

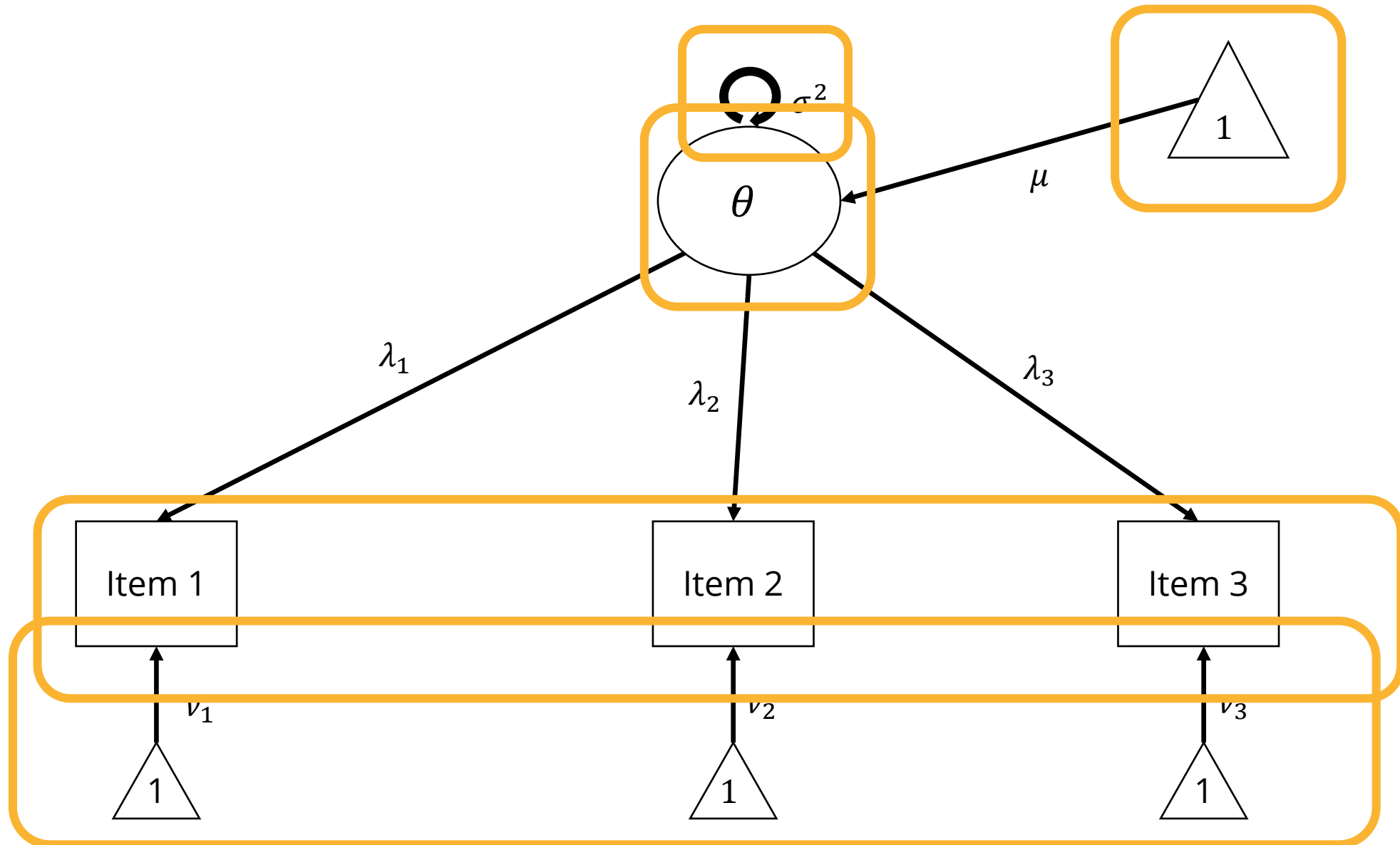
Multigroup IRT-based methods

- Based on IRT parameterization
- **Allows for simultaneous modeling of DIF and mean/variance impact**
- Categorical background variables only
- More than one grouping variable? Must be combined to create mutually exclusive groups

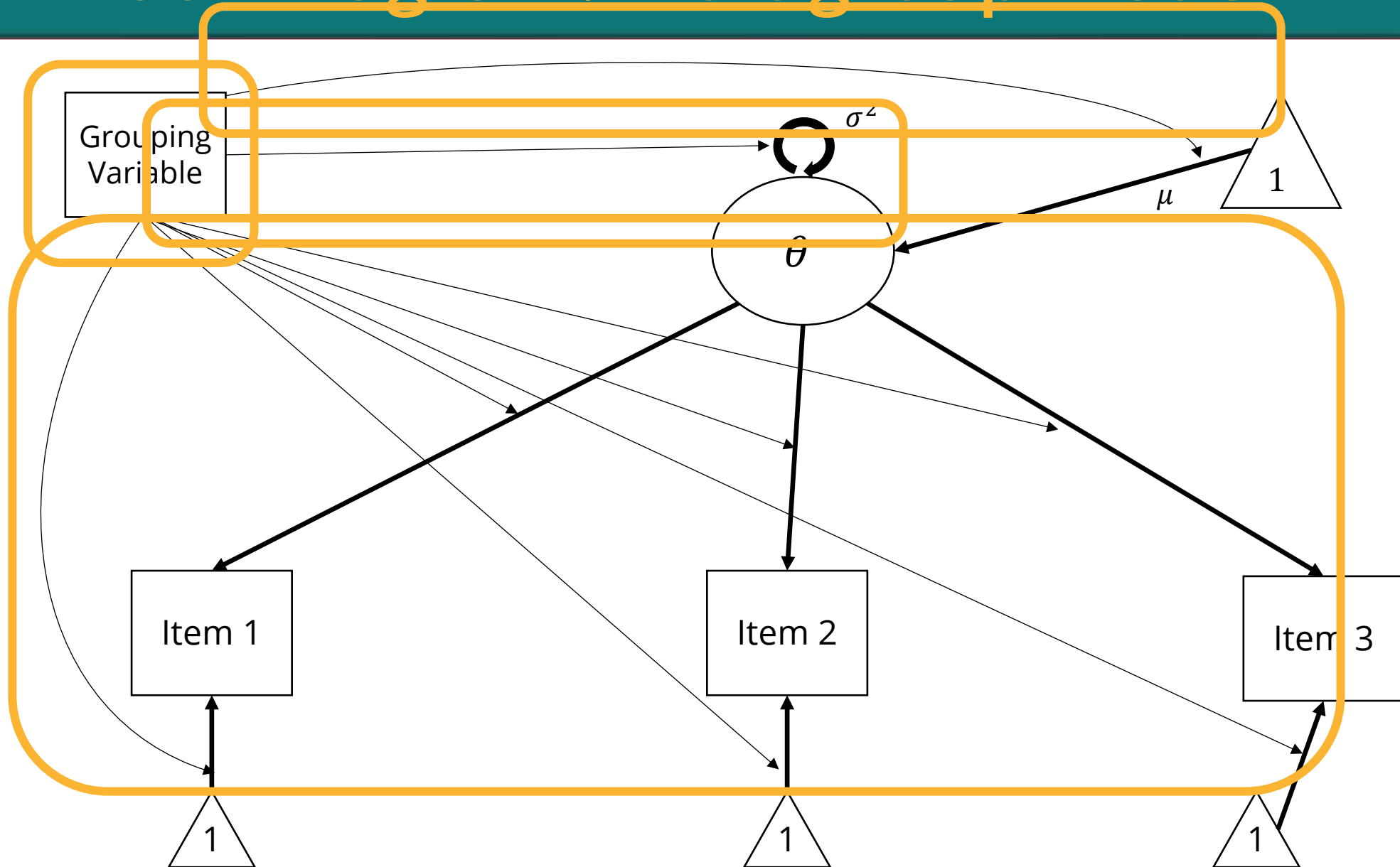
SEM-based MIMIC-with-interaction models

- **Based on slope-intercept parameterization**
- Allows for simultaneous modeling of DIF and mean impact
- **Categorical, continuous, nominal, censored, etc. background variables; groupings can overlap**

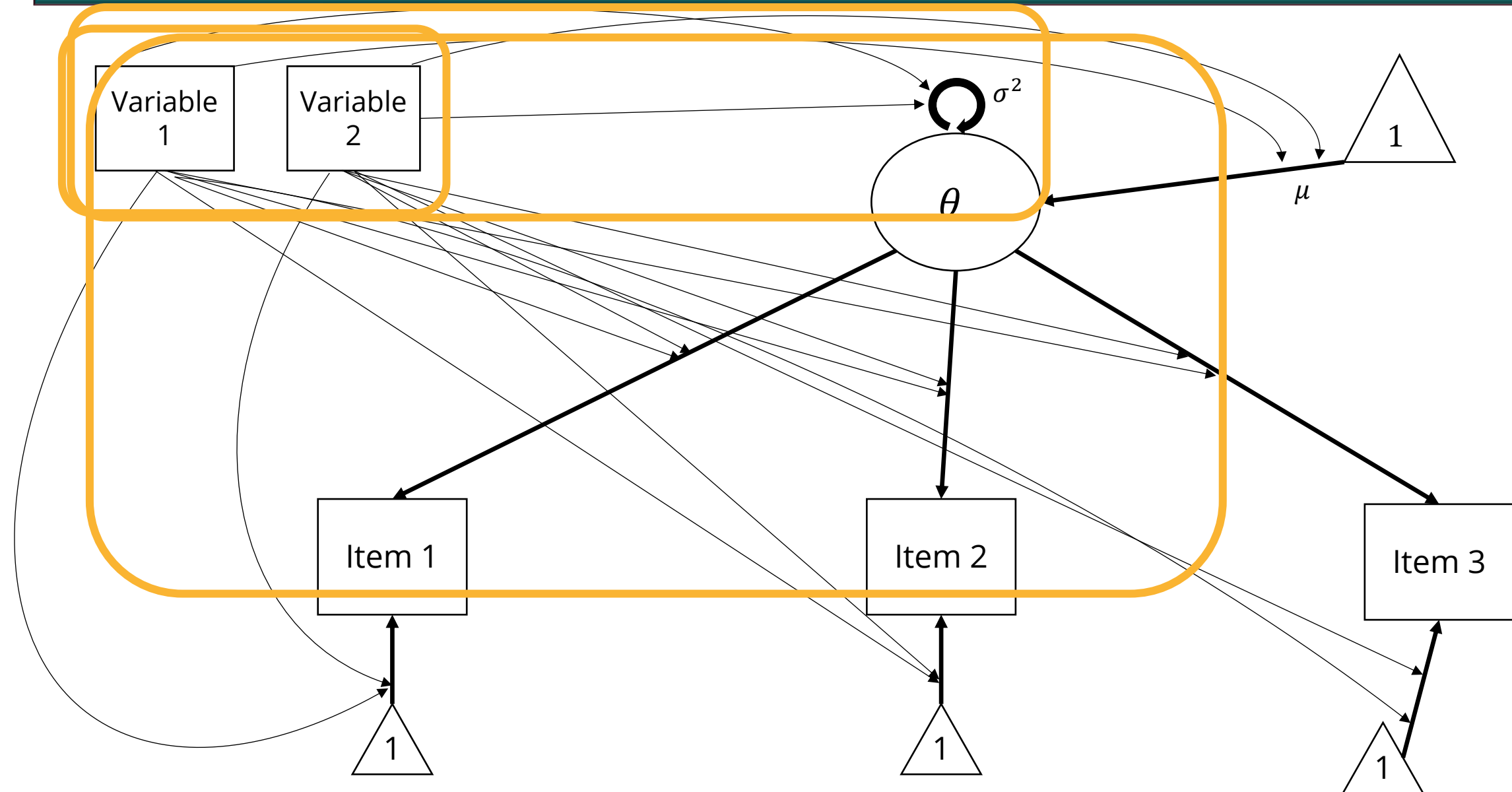
Path Diagram: 2PL



Path Diagram: Multigroup Model



Path Diagram: MNLFA



Section 1 References

- Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, 22(3), 507–526. <https://doi.org/10.1037/met0000077>
- Bauer, D. J. (2023). Enhancing measurement validity in diverse populations: Modern approaches to evaluating differential item functioning. *British Journal of Mathematical and Statistical Psychology*, 76(3), 435–461. <https://doi.org/10.1111/bmsp.12316>
- Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, 14(2), 101–125. <https://doi.org/10.1037/a0015583>
- Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. L. Erlbaum Associates.
- Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and Item Response Theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, 15(1), 136–153. <https://doi.org/10.1080/10705510701758406>
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525–543. <https://doi.org/10.1007/BF02294825>
- Russell, M. (2024). Digital module 36: Applying intersectionality theory to educational measurement. *Educational Measurement: Issues and Practice*, 43(3), 106–108. <https://doi.org/10.1111/emip.12622>
- Thissen, D. (2024). A review of some of the history of factorial invariance and differential item functioning. *Multivariate Behavioral Research*. Advance online publication. <https://doi.org/10.1080/00273171.2024.2396148>
- Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 147–172). Lawrence Erlbaum Associates, Inc.
- Xu, R., & Soland, J. (2024). Beyond group comparisons: Accounting for intersectional sources of bias in international survey measures. *International Journal of Testing*, 24(3), 230–258. <https://doi.org/10.1080/15305058.2024.2364168>

An Overview of MNLFA for DIF and Impact Assessment

2

Section Learning Objectives

2

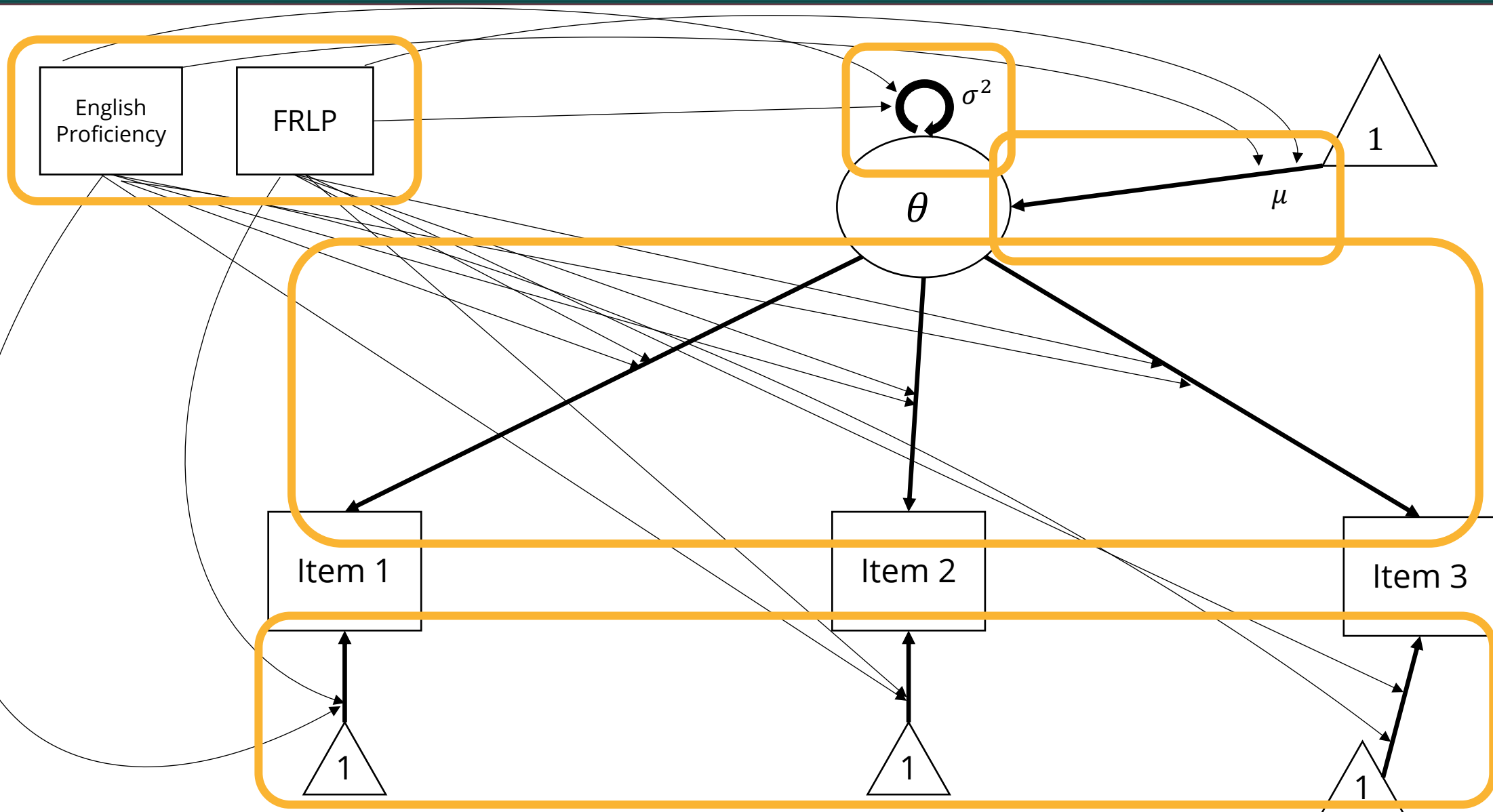
MNLFA Overview

Recognize and describe the paths in an MNLFA path diagram

Connect the path diagram to DIF via the MNLFA item response function

Connect the path diagram to the MNLFA representation of impact

Path Diagram: MNLFA

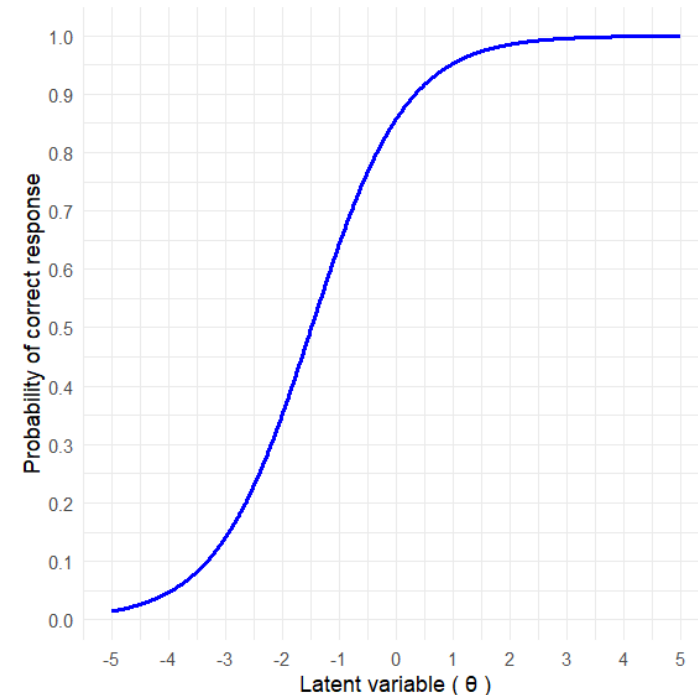


The Slope-Intercept 2PL

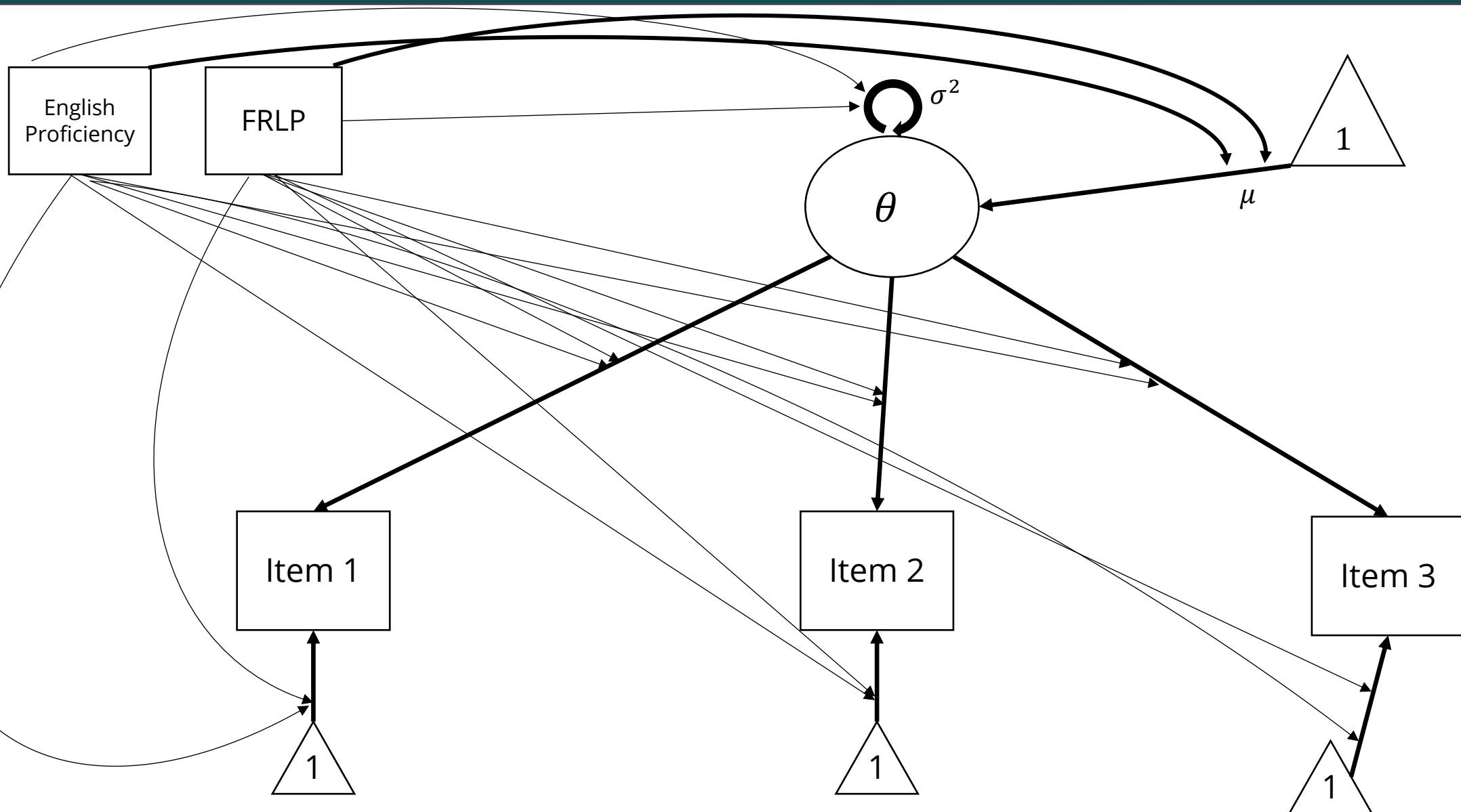
$$P(X_{ij} = 1) = \frac{\exp(\lambda_j \theta_i + \tau_j)}{1 + \exp(\lambda_j \theta_i + \tau_j)}, \theta \sim N(\mu, \sigma^2)$$

$$\ln \left(\frac{P(X_{ij} = 1)}{1 - P(X_{ij} = 1)} \right) = \text{logit}(X_{ij} = 1) = \lambda_j \theta_i + v_j$$

- $\lambda_j = a_j$
- $v_j = -b_j a_j$
- Let $\lambda_i = 1.2$
- Let $v_j = 1.8$
 - ($b_j = -1.5$)



Path Diagram: Mean Impact



Mean Impact

$$\theta \sim N(\mu, \sigma^2)$$

- Mean impact is modeled as moderation of μ by covariates:

$$\theta \sim N(\mu + \alpha' \mathbf{z}_i, \sigma^2)$$

- For our two background variables:

$$\theta \sim N(\mu + \alpha_1 z_{1i} + \alpha_2 z_{2i}, \sigma^2)$$

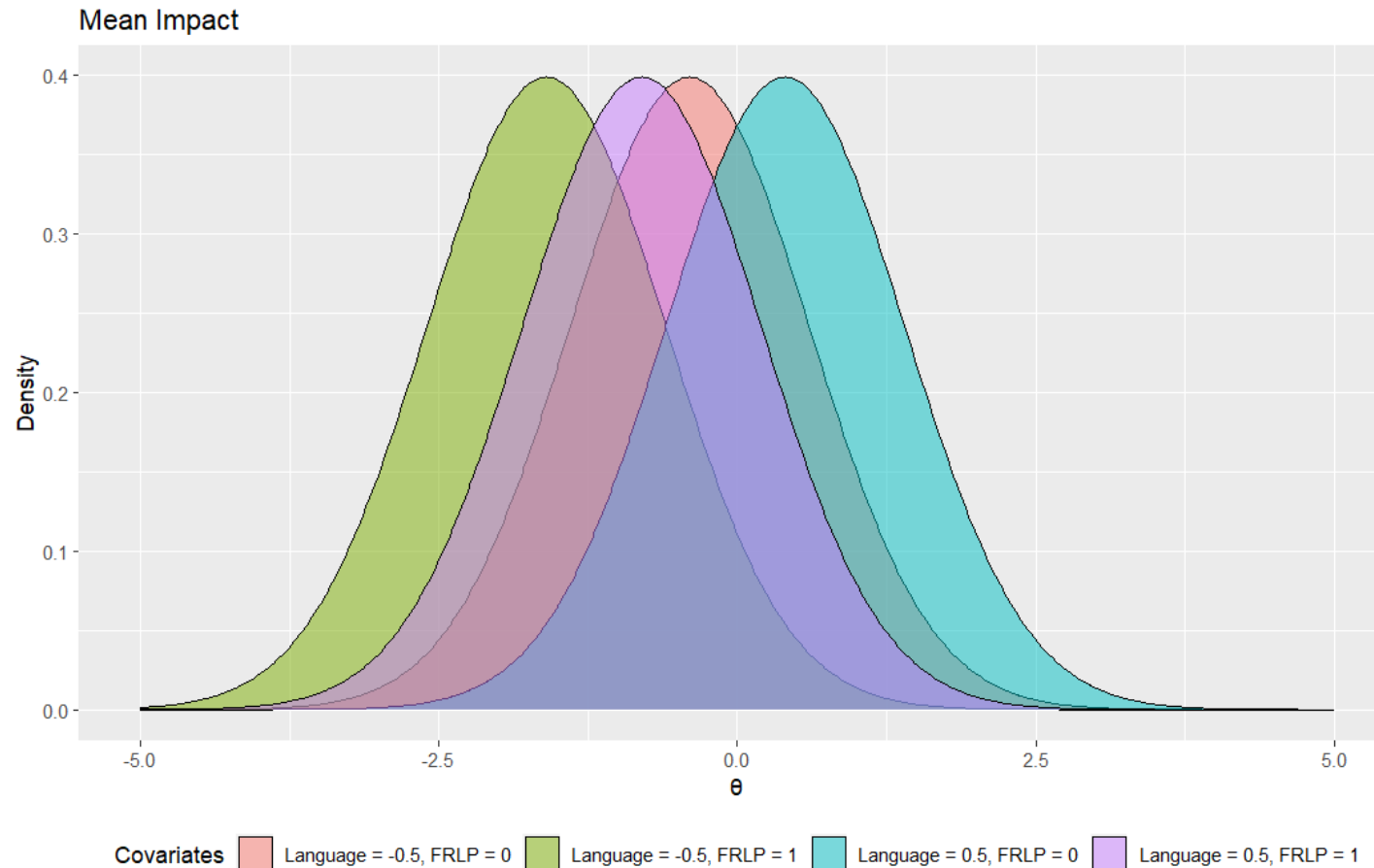
- Commonly,

$$\theta \sim N(0 + \alpha_1 z_{1i} + \alpha_2 z_{2i}, 1)$$

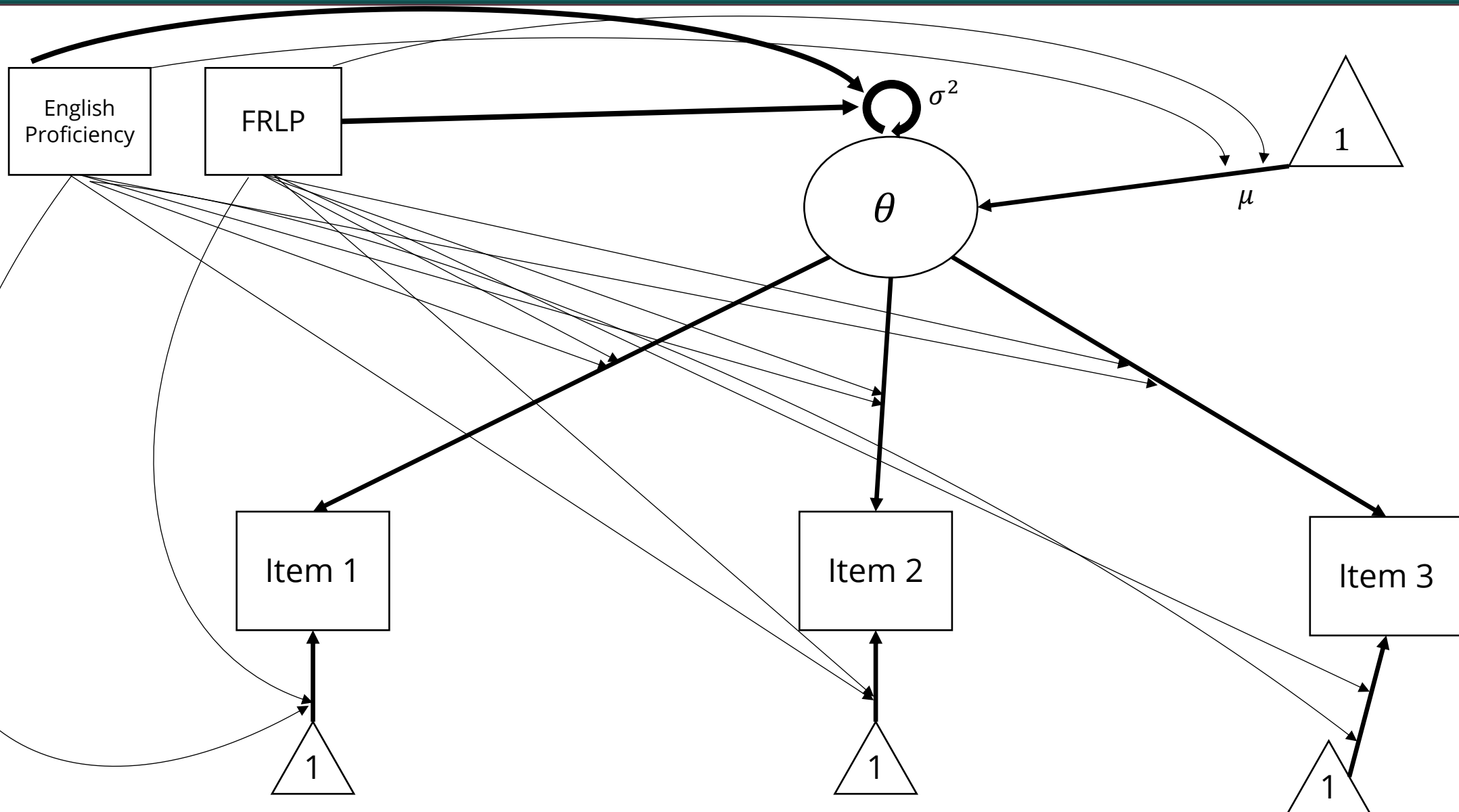
Mean Impact

$$\theta \sim N(\mu + \alpha_1 \text{lang}_i + \alpha_2 \text{FRLP}_i, \sigma^2)$$

- Let $\alpha_1 = 0.8, \alpha_2 = -1.2$



Path Diagram: Variance Impact



Variance Impact

$$\theta \sim N(\mu + \alpha' \mathbf{z}_i, \sigma^2)$$

- In addition to moderation of μ , MNLFA also allows for moderation of σ^2 :

$$\theta \sim N(\mu + \alpha' \mathbf{z}_i, \sigma^2 * e^{\omega' \mathbf{z}_i})$$

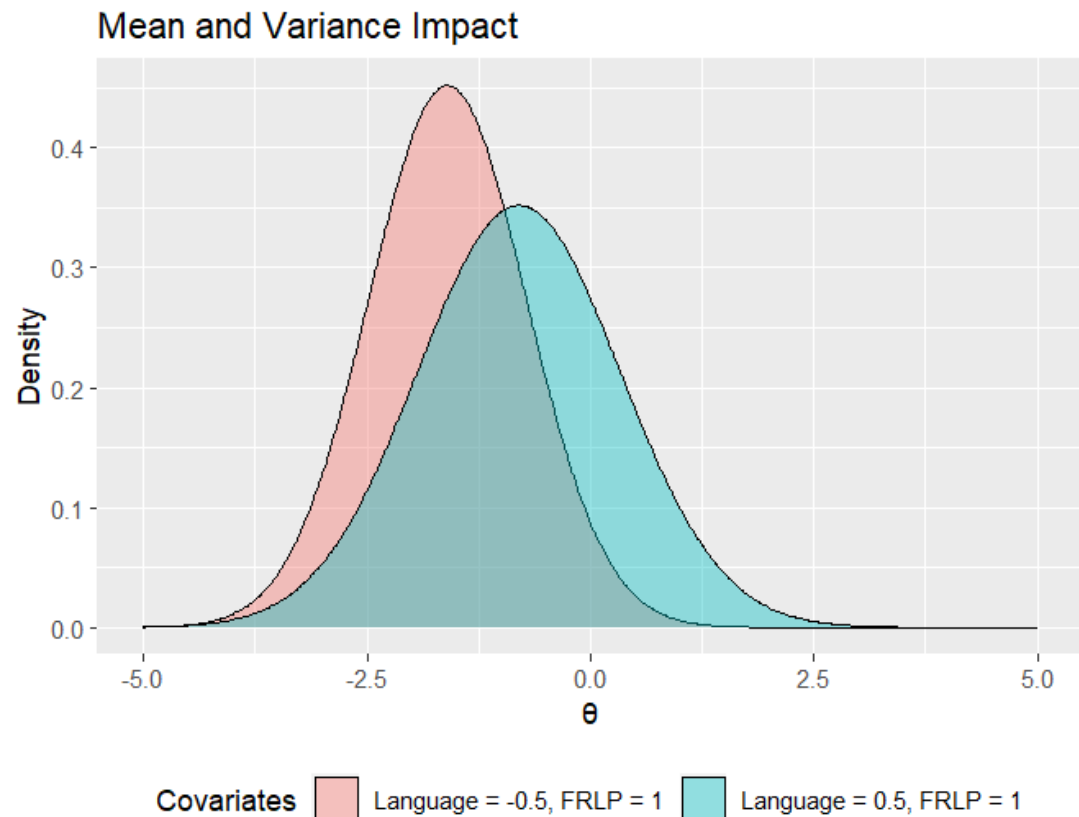
- For our two background variables:

$$\theta \sim N(\mu + \alpha_1 z_{1i} + \alpha_2 z_{2i}, \sigma^2 * e^{\omega_1 z_{1i} + \omega_2 z_{2i}})$$

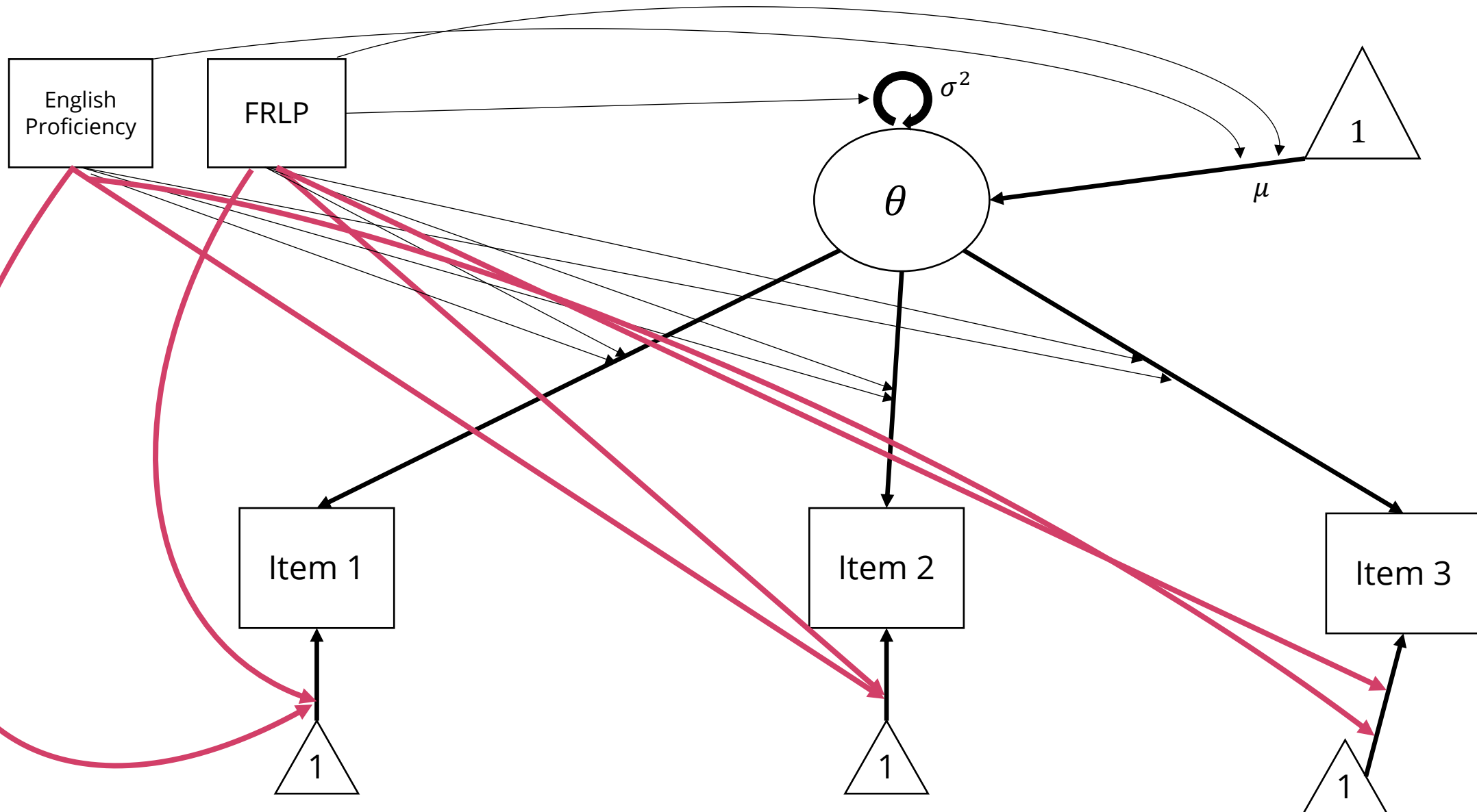
Mean and Variance Impact

$$\theta \sim N(\alpha_1 \text{lang}_i + \alpha_2 \text{FRLP}_i, e^{\omega_1 \text{lang}_i + \omega_2 \text{FRLP}_i})$$

- For identification,
 $\mu = 0, \quad \sigma^2 = 1$
- $\alpha_1 = 0.8, \alpha_2 = -1.2, \omega_1 = 0.5, \omega_2 = 0$
 - *FRLP* has no impact on variance
 - Plot: *lang* score of -0.5, 0.5; *FRLP* is 1 for both distributions



Path Diagram: Intercept/Uniform DIF



Uniform DIF

$$\text{logit}(X_{ij} = 1) = \lambda_j \theta_i + \nu_j + \mathbf{v}_j' \mathbf{z}_i, \quad \theta \sim N(\mu, \sigma^2)$$

- This equation adds in uniform DIF to the slope-intercept model we presented a couple of slides ago.
- i indexes a person, j indexes an item
- ν_j = item intercept, λ_j = item slope
- \mathbf{v}_j' is a vector of uniform DIF parameters for item j (transposed)
- \mathbf{z}_i is a vector of background variable values for person i

Uniform DIF with Two Variables

$$\text{logit}(X_{ij} = 1) = \lambda_j \theta_i + \nu_j + \mathbf{v}_j' \mathbf{z}_i, \quad \theta \sim N(\mu, \sigma^2) \Rightarrow$$

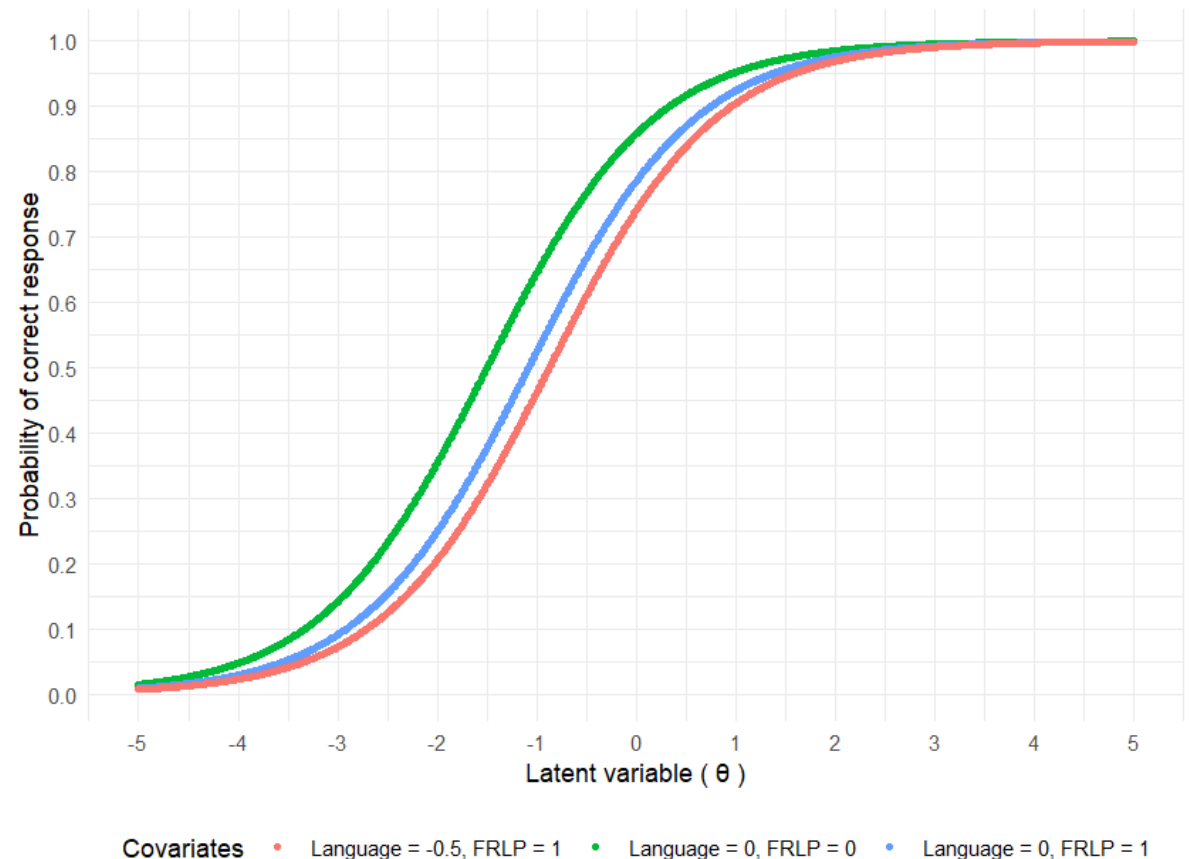
- Two background variables – we'll replace the vectors with the two parameters that represent uniform DIF.

$$\text{logit}(X_{ij} = 1) = \lambda_j \theta_i + \nu_j + \nu_{1j} z_{1i} + \nu_{2j} z_{2i}, \quad \theta \sim N(\mu, \sigma^2)$$

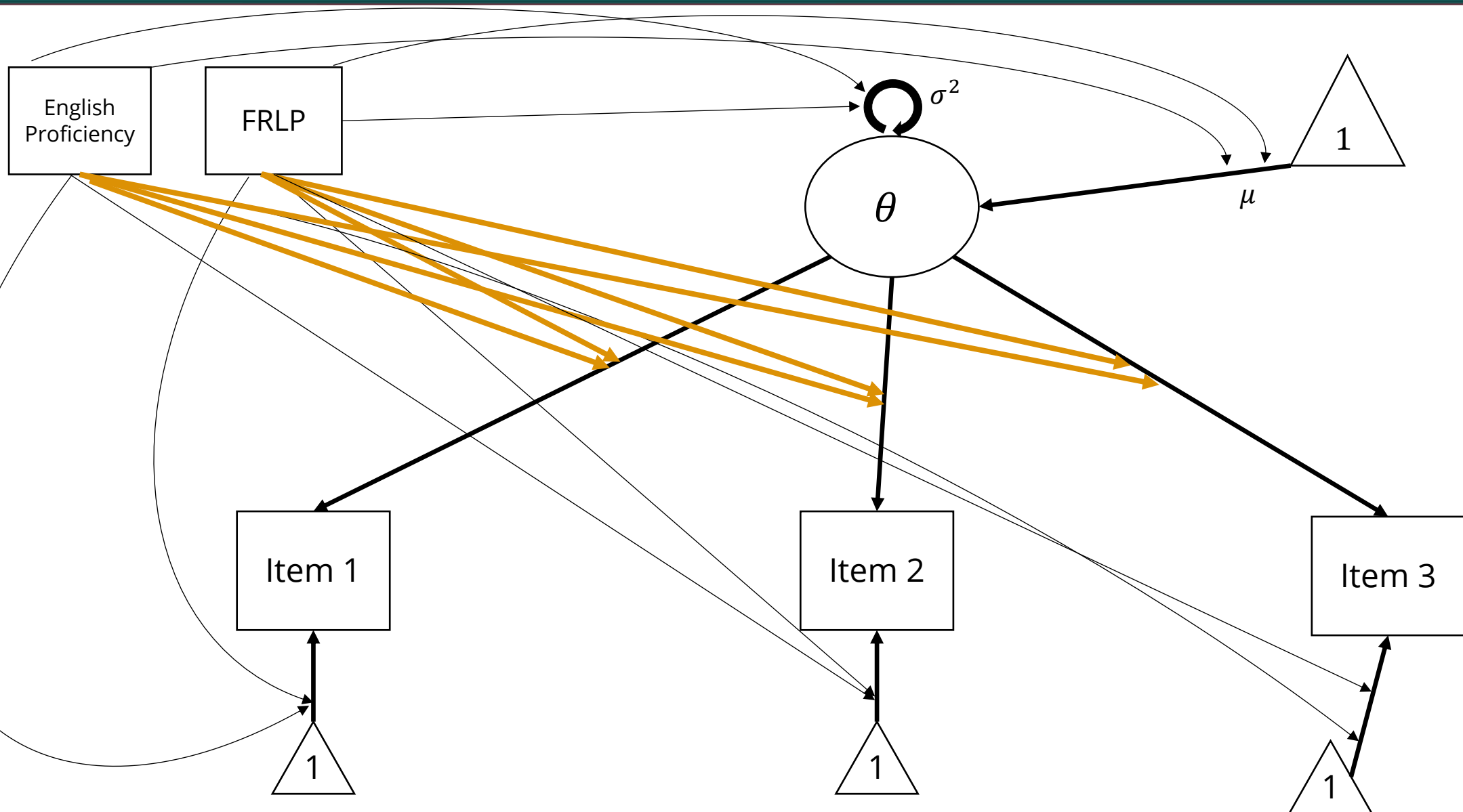
Uniform DIF with Two Variables

$$\text{logit}(X_{ij} = 1) = \lambda_j \theta_i + v_j + 0.5 \text{lang}_i - 0.5 \text{FRLP}_i$$

- Let's say that:
 - lang_i is person i 's z-scored language proficiency
 - FRLP_i is that person's FRLP status (0 = ineligible, 1 = eligible)



Path Diagram: Nonuniform DIF



Nonuniform DIF

$$\text{logit}(X_{ij} = 1) = (\lambda_j + \lambda_j' \mathbf{z}_i) \theta_i + \nu_j + \nu_j' \mathbf{z}_i$$

- This equation adds in nonuniform DIF to the model.
- λ_j is a vector of nonuniform DIF parameters (moderation of slope)
- Each value in λ_j increases or decreases the item's discrimination according to a background variable

Nonuniform DIF with Two Variables

$$\text{logit}(X_{ij} = 1) = (\lambda_j + \boldsymbol{\lambda}_j' \mathbf{z}_i) \theta_i + \nu_j + \boldsymbol{\nu}_j' \mathbf{z}_i, \quad \theta_i \sim N(\mu, \sigma^2)$$

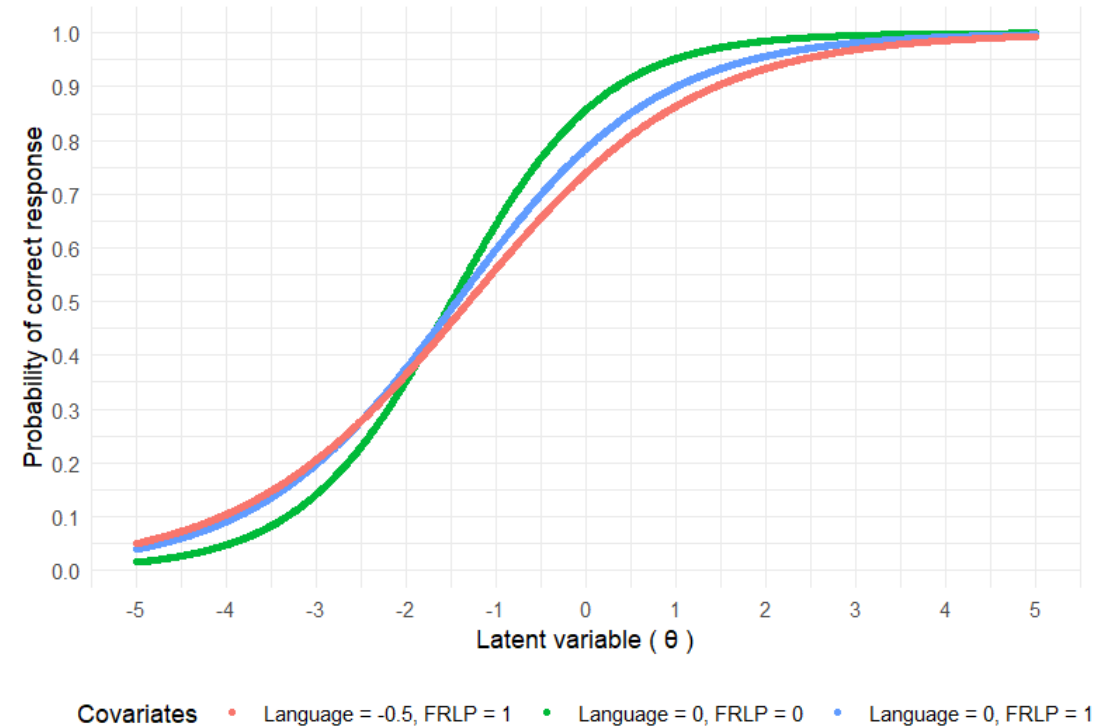
- Now, we'll be replacing the vectors with parameters that represent uniform and nonuniform DIF.

$$\text{logit}(X_{ij} = 1) = (\lambda_j + \lambda_{1j} z_{1i} + \lambda_{2j} z_{2i}) \theta_i + \nu_j + \nu_{1j} z_{1i} + \nu_{2j} z_{2i}$$

Nonuniform DIF with Two Variables

$$\begin{aligned} \text{logit}(X_{ij} = 1) \\ = (\lambda_j + 0.2\text{lang}_i - 0.3\text{FRLP}_i)\theta_i + v_j + 0.5\text{lang}_i - 0.5\text{FRLP}_i \end{aligned}$$

- In addition to the uniform DIF we already added in:
 - $\text{lang} = +0.2$ to slope (more discriminating)
 - $\text{FRLP} = -0.3$ to slope (less discriminating)



The Full Model

$$\text{logit}(X_{ij} = 1) = (\lambda_j + \lambda_j' \mathbf{z}_i) \theta_i + \nu_j + \nu_j' \mathbf{z}_i,$$

$$\theta \sim N(\mu + \alpha' \mathbf{z}_i, \sigma^2 * e^{\omega' \mathbf{z}_i})$$

- $\lambda_j' \mathbf{z}_i$: Moderation of slope (discrimination)
- $\nu_j' \mathbf{z}_i$: Moderation of intercept (difficulty)
- $\alpha' \mathbf{z}_i$: Moderation of mean (mean impact)
- $e^{\omega' \mathbf{z}_i}$: Moderation of variance (variance impact)
- Everything else: parameters of slope-intercept 2PL

References

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, 22(3), 507–526. <https://doi.org/10.1037/met0000077>

MNLFA Estimation and Interpretation

3

3 MNLFA Estimation and Interpretation

Section Learning Objectives

Describe options for estimating an MNLFA's parameters

Interpret the parameters of an MNLFA analogous to regression

Describe score estimation in the presence of DIF and impact

Interpret θ and model fit in presence of DIF and impact

MNLFA Estimation

- MNLFAs are more challenging models to estimate than non-moderated factor/IRT models due to added parameters of background variables
- Maximum Likelihood (Bauer, 2017)
 - Typically specified through non-linear constraints on specific parameters
 - Theoretically-specified moderated relations
 - Allows the greater user-level control
 - Intensive estimation because all moderation parameters are freely estimated
 - A dimensionality problem!

MNLFA Estimation (Cont.)

- Regularization (Belzak & Bauer, 2020; 2024)
 - Penalized covariate effects on targeted parameters – automatic variable selection
 - Allows for a greater number of potential moderated effects
 - Selection of regularization approach impacts results
 - Ridge, LASSO, Elastic Net, etc.
 - Lasso will tend to select one large relationship and shrink others towards zero
 - Ridge will tend to shrink the effects of groups of correlated covariates towards one another

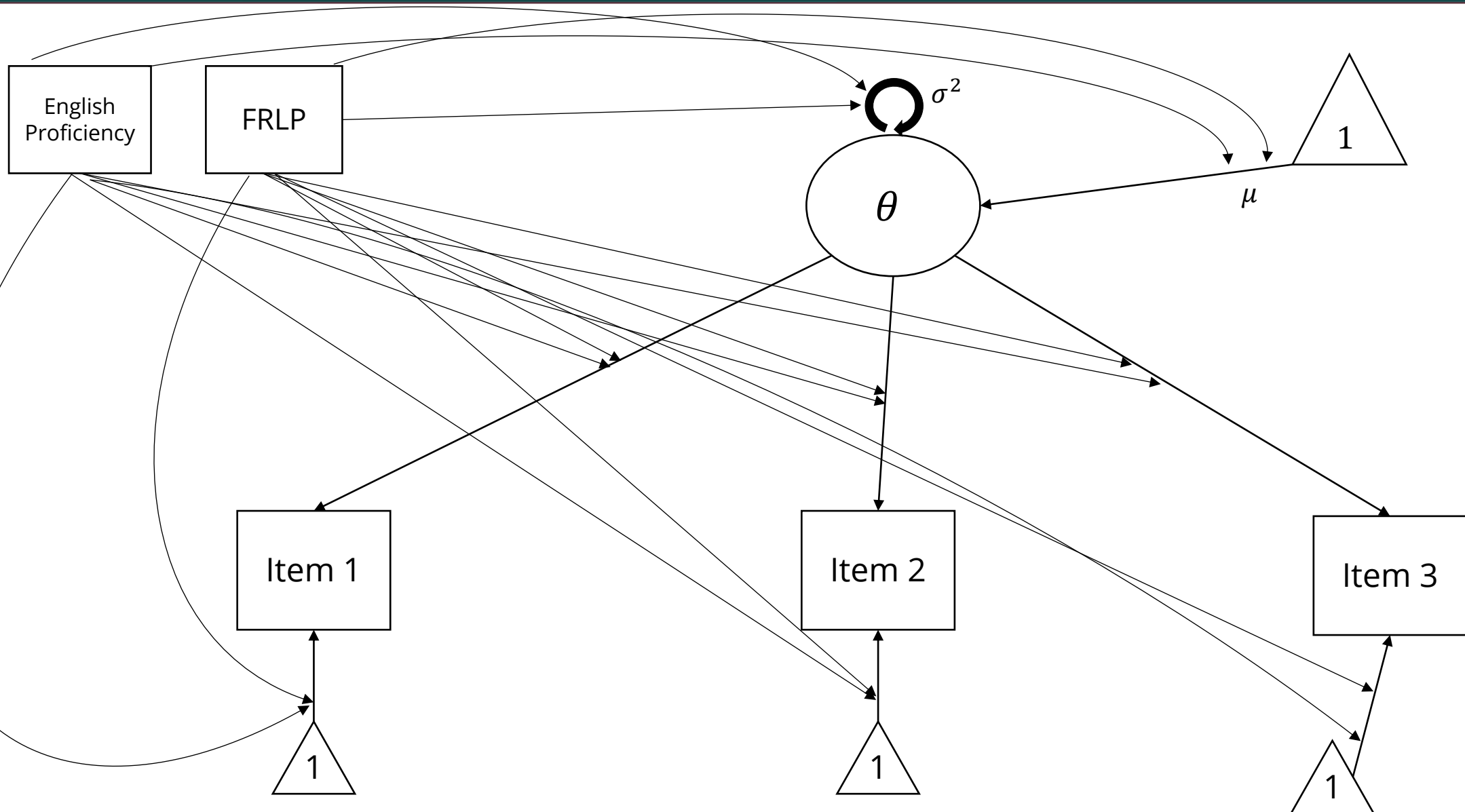
MNLFA Estimation (Cont.)

- Bayesian estimation (Chen et al., 2022; Brandt et al., 2023; Enders et al., 2024)
 - Parameter moderation is naturally accommodated in general Bayesian framework; regularization accomplished via priors (e.g., Laplace, Spike-and-Slab, Horseshoe priors; Spike-and-slab seem to work best)
 - Non-regularized versions also available for estimation advantages

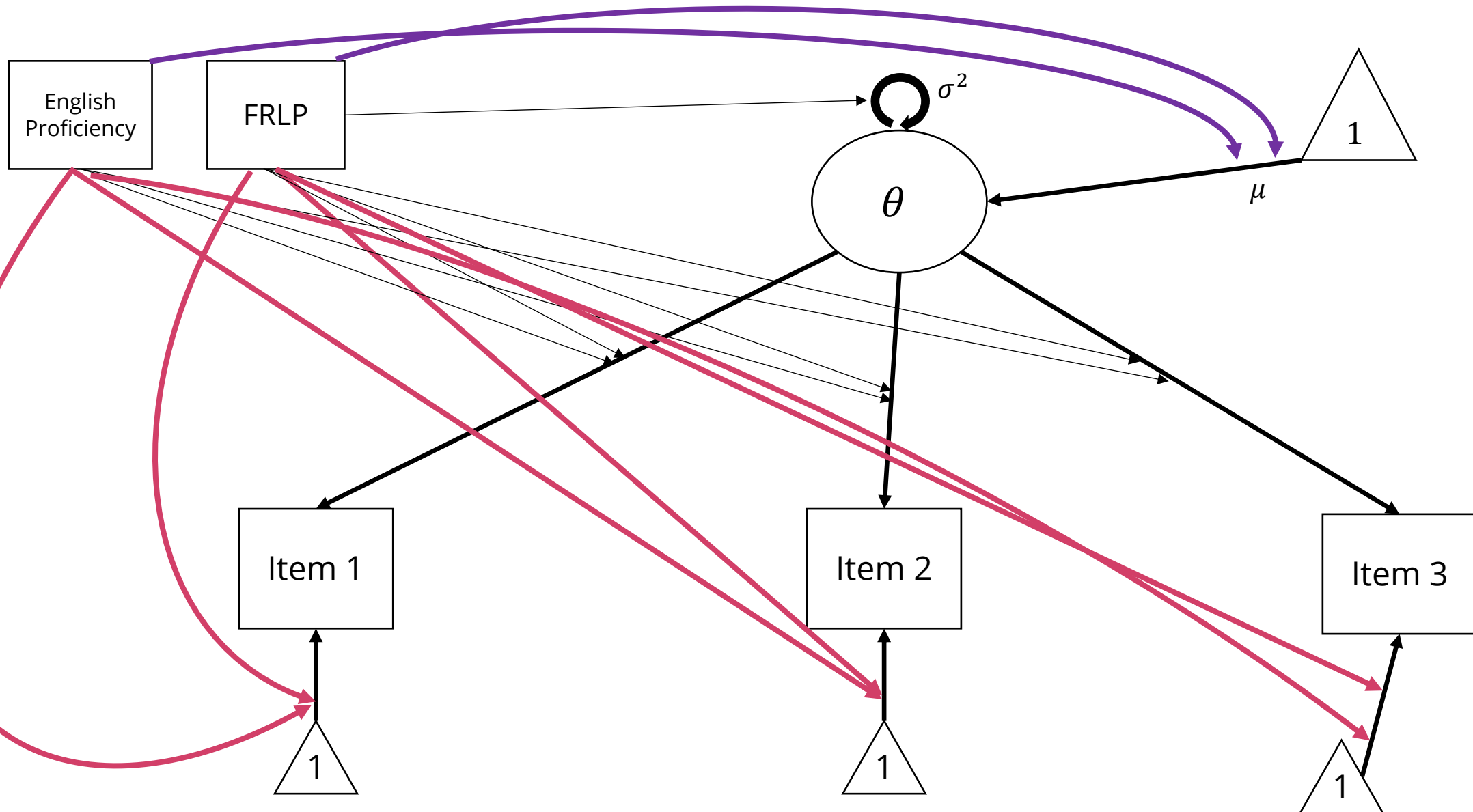
Back to the Model...

No matter how the model is estimated, we ultimately end up interpreting the same kinds of parameters.

Path Diagram: MNLFA



Interpreting Intercept Moderation



Interpreting Intercept Moderation

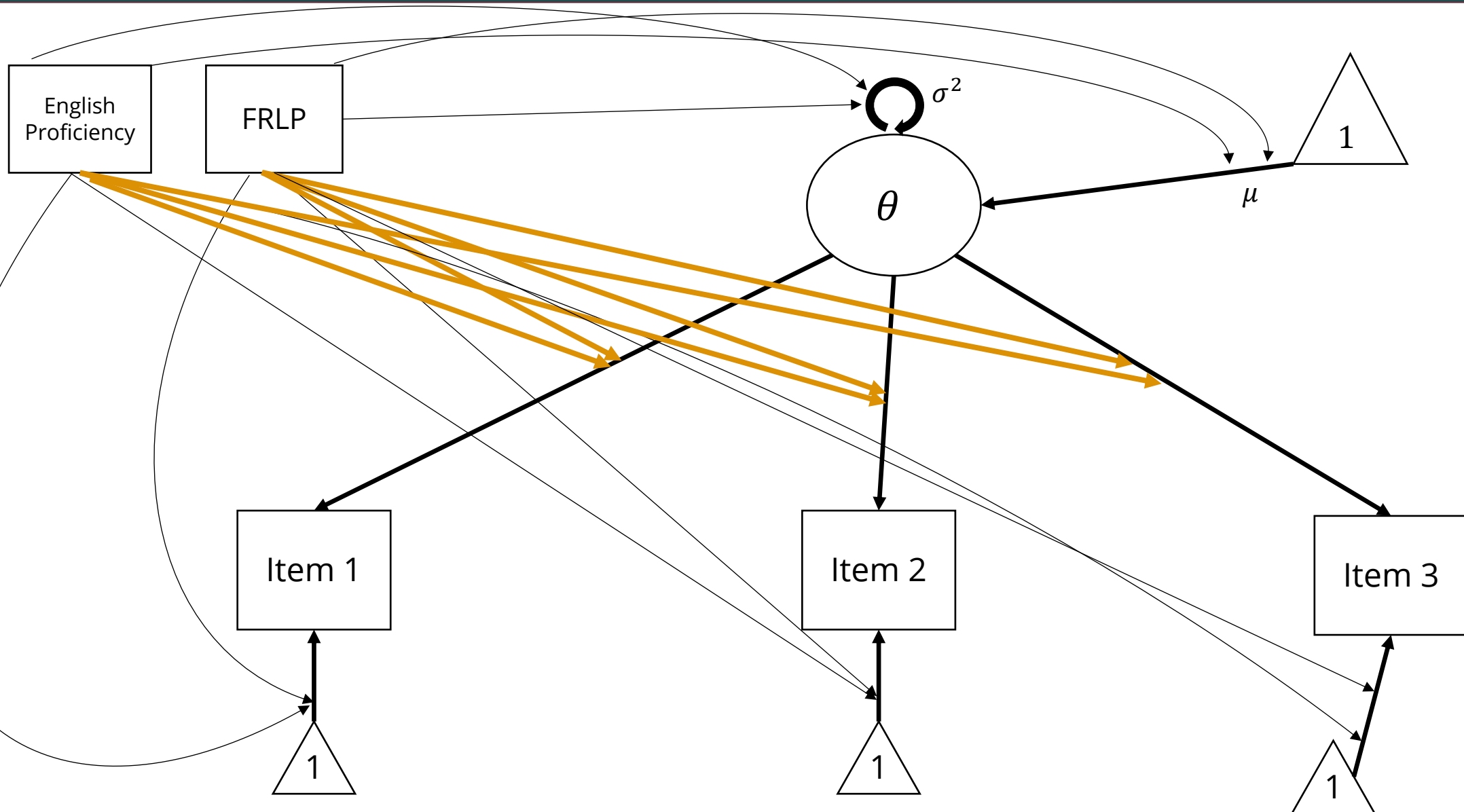
$$\text{logit}(X_{ij} = 1) = \lambda_j \theta_i + v_j + \mathbf{0.5lang}_i - 0.5FRLP_i$$

- Moderation of item intercept: a one-unit change in language score is associated with the item becoming **0.5 logits easier (uniform DIF)**

$$\theta \sim N(\mu + \mathbf{0.8lang}_i - 1.2FRLP_i, \sigma^2)$$

- Moderation of θ intercept: a one-unit change in language score is associated with average θ being **0.8 units higher (mean impact)**

Interpreting Slope Moderation

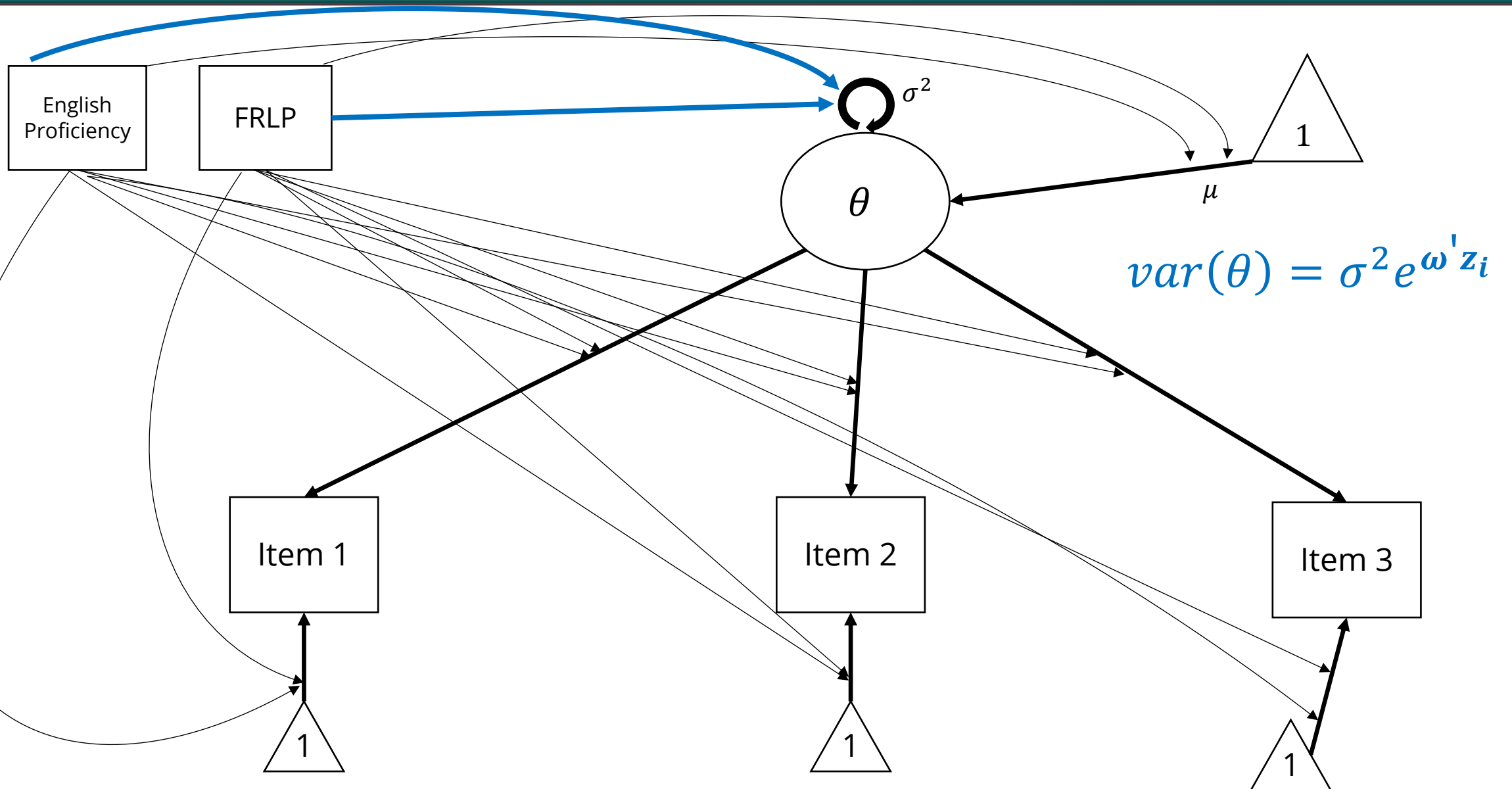


Interpreting Slope Moderation

$$\begin{aligned} \text{logit}(X_{ij} = 1) \\ = (\lambda_j + \mathbf{0.2lang_i} - 0.3FRLP_i)\theta_i + v_j + 0.5lang_i - 0.5FRLP_i \end{aligned}$$

- Moderation of item slope: a one-unit change in language score is associated with a **0.2 unit increase in the item slope (nonuniform DIF)**, i.e., the item is *more discriminating* for examinees with higher language scores

Path Diagram: Variance Impact

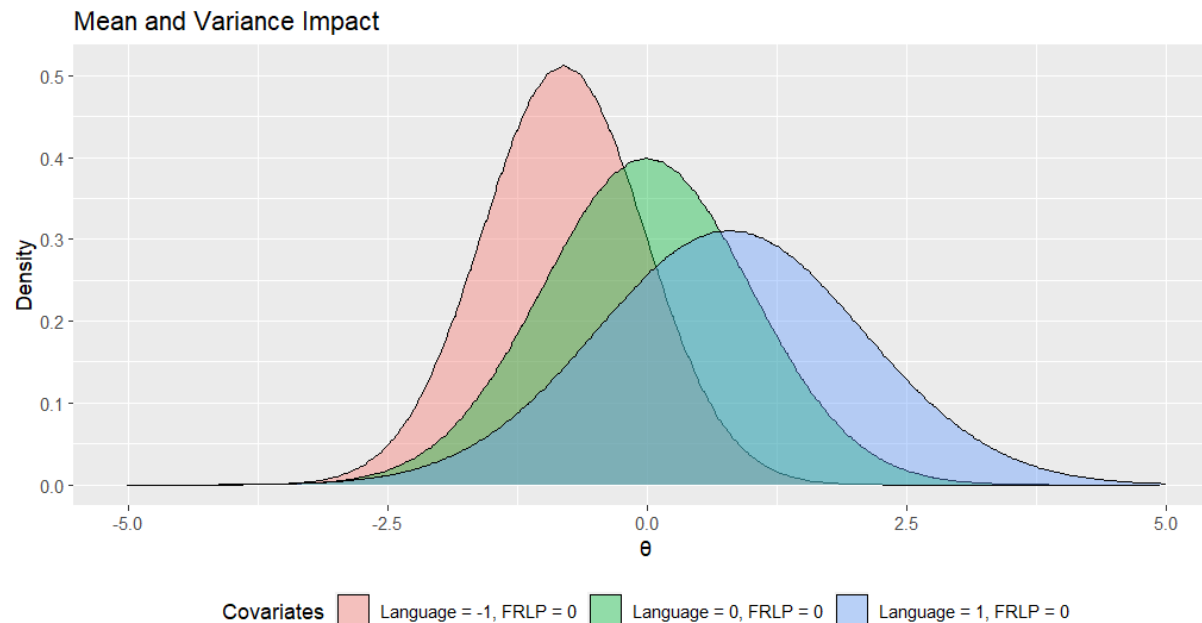


Interpreting Variance Moderation

$$\theta \sim N(0 + 0.8\text{lang}_i + 1.2\text{FRLP}_i, 1 * e^{0.5\text{lang}_i + 0 * \text{FRLP}_i})$$

- Moderation of variance: a one-unit change in language score is associated with a 0.5 unit increase in the log of the variance (i.e., higher language scores -> more score variance)

- $e^{0.5 * -1} = 0.61$
- $e^{0.5 * 0} = 1$
- $e^{0.5 * 1} = 1.65$



Treatment of DIF Items

- While often items showing DIF are removed, this is not always possible or desirable (Curran et al., 2016)
 - Specific background knowledge on tests likely persist
 - As a means to understand how different test takers interpret or react to different information, not just an investigation of potentially unfair test items
 - Differential expression of latent trait may be of theoretical interest or point to policy interventions
 - May result in too few items in instrument
 - Scoring is possible in the presence of covariate effects; whether to drop or retain DIF items depends on application and background variables

Estimation of Scores

- Major approaches include:
 - Expected A Posteriori (EAP) or Maximum A Posteriori (MAP) scores can be computed that incorporate covariate information
 - Multi-step scoring process using a calibrated sample to generate parameter estimate is recommended in complex samples (e.g., longitudinal, nested data; Bauer et al., 2020)

Interpretation of θ and Fit

- Assuming proper DIF/impact effect specification, θ can be interpreted as usual
 - Mis-specification (as always) will bias θ whether scores are obtained or not
 - Tend towards maximal specification of covariate effects
 - Regularization approaches will introduce bias

Interpretation of θ and Fit

- Null-model based fit indices do not apply to MNLFA
 - What is the correct baseline model?
- Many simpler models are nested within the MNLFA, allowing likelihood ratio model comparison tests for fit
 - Standard IRT with no moderation, multiple-groups models
 - Information criterion (AIC, BIC, aBIC) comparisons also possible

Further Topics

- Polytomous, continuous items
 - Threshold vs. intercept DIF
 - With ordered categorical items, the number of parameters estimated can become incredibly large
 - Estimate one intercept DIF parameter per background variable per item
- Power
 - Due to the complexity of the models, it can be difficult to conduct a Monte Carlo simulation to detect power
 - Challenging to assess with many background variables and items

Further Topics (Cont.)

- Asymmetry of DIF
 - DIF may be easier to detect when it is symmetric (Halpin, 2024; DeMars, 2020)
 - When there are DIF items favoring the groups being compared at roughly equal rates
- Intersectionality
 - Interactions between background variables

References

- Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, 22(3), 507–526. <https://doi.org/10.1037/met0000077>
- Bauer, D. J., Belzak, W. C. M., & Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(1), 43–55. <https://doi.org/10.1080/10705511.2019.1642754>
- Belzak, W. C. M., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, 25(6), 673–690. <https://doi.org/10.1037/met0000253>
- Belzak, W. C. M., & Bauer, D. J. (2024). Using regularization to identify measurement bias across multiple background characteristics: A penalized expectation–maximization algorithm. *Journal of Educational and Behavioral Statistics*. Advance online publication. <https://doi.org/10.3102/10769986231226439>
- Brandt, H., Chen, S. M., & Bauer, D. J. (2023). Bayesian penalty methods for evaluating measurement invariance in moderated nonlinear factor analysis. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000552>
- Chen, S. M., Bauer, D. J., Belzak, W. M., & Brandt, H. (2022). Advantages of spike and slab priors for detecting differential item functioning relative to other Bayesian regularizing priors and frequentist lasso. *Structural Equation Modeling: A Multidisciplinary Journal*, 29(1), 122–139. <https://doi.org/10.1080/10705511.2021.1948335>
- Curran, P. J., Cole, V., Bauer, D. J., Hussong, A. M., & Gottfredson, N. (2016). Improving factor score estimation through the use of observed background characteristics. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(6), 827–844. <https://doi.org/10.1080/10705511.2016.1220839>
- DeMars, C. E. (2020). Alignment as an alternative to anchor purification in DIF analyses. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(1), 56–72. <https://doi.org/10.1080/10705511.2019.1617151>
- Enders, C. K., Vera, J. D., Keller, B. T., Lenartowicz, A., & Loo, S. K. (2024). Building a simpler moderated nonlinear factor analysis model with Markov Chain Monte Carlo estimation. *Psychological Methods*. <https://doi.org/10.1037/met0000712>
- Halpin, P. F. (2024). Differential item functioning via robust scaling. *Psychometrika*. Advance online publication. <https://doi.org/10.1007/s11336-024-09957-6>

MNLFA Applied Example

4

4

MNLFA Applied Example

Section Learning Objectives

Summarize data and model

Detail regularization approach used for
example model estimation

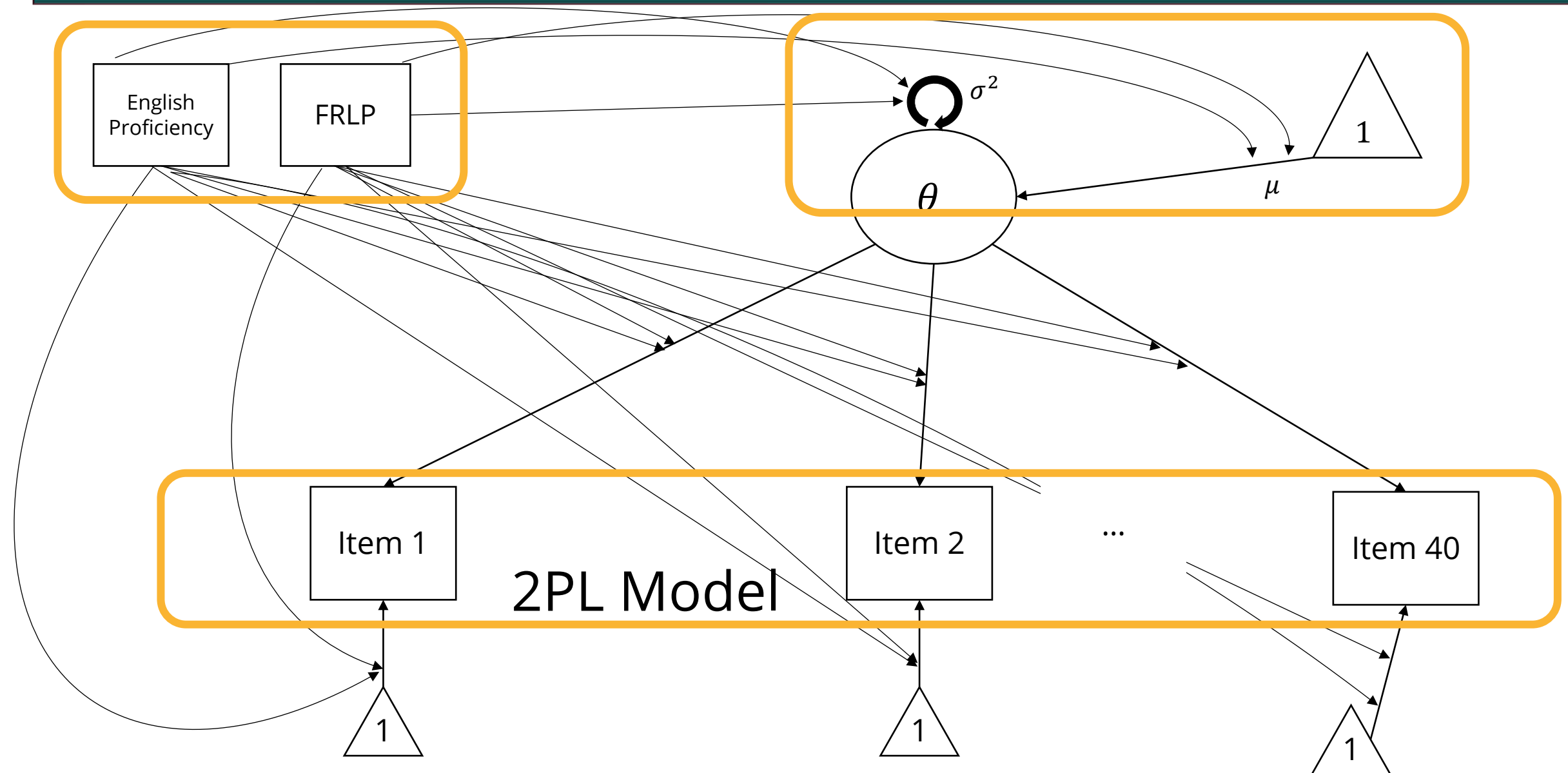
Generate and interpret parameter
estimates

Recognize and conduct important next
steps

Applied Example: Data

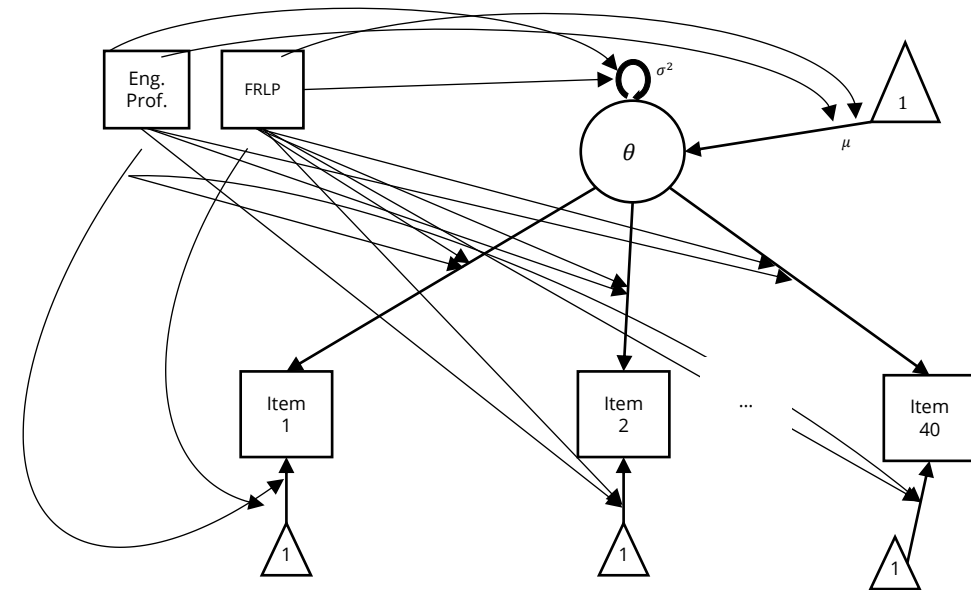
- Simulated dataset
- $N = 10,000$ students
- 40 binary items (1 = correct) from a math ability test
- Binary SES variable based on free-or-reduced-price lunch eligibility (coded such that 0 = eligible/lower SES, 1 = ineligible/higher SES)
- Gaussian IRT-derived score of English proficiency ($\mu_{\theta_l} = 0, \sigma_{\theta_l}^2 = 1$)

Applied Example: Model



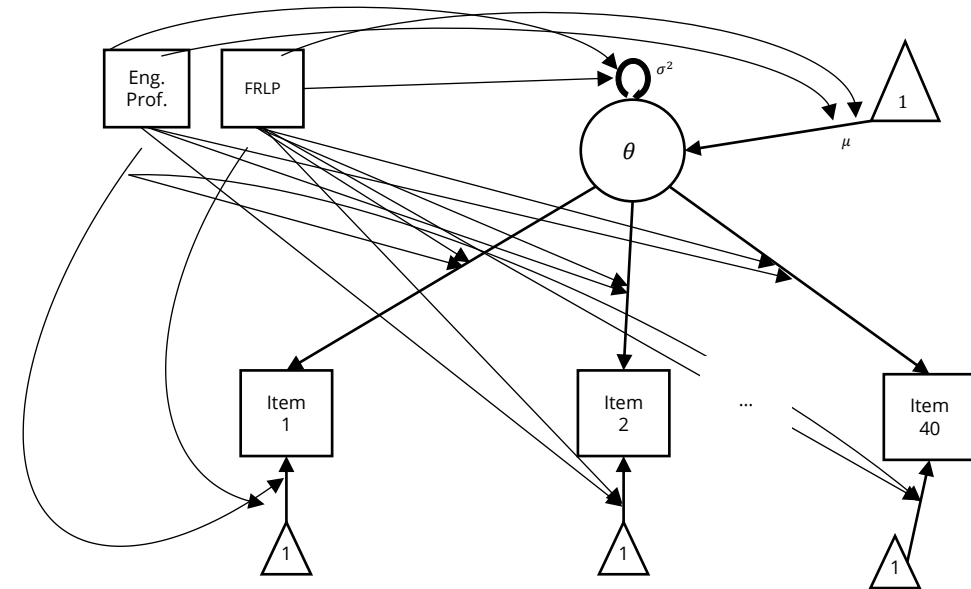
Applied Example: Model Estimation

- R package `{regDIF}`
 - Automatically specifies all DIF and impact relations
 - Uses regularization to prune DIF effects, but does not penalize impact (Bauer et al., 2019)
- Will use LASSO penalty
 - Others available (Ridge, elastic net), see Belzak (2023)
 - Optimal regularization selected using BIC
 - 100 tuning parameter values evaluated



Applied Example: Model Estimation

- A note on background variables
 - *regDIF* standardizes all variables by default – necessary for regularization to work properly
 - This means that categorical variables will be effect-coded
 - Affects interpretation – but generally not an issue because model will be re-estimated



Applied Example: Results

- Identifying non-zero DIF effects on intercept and slope parameters

FRLP	
Param.	Regularized Effect
ν_2	0.015
ν_3	0.303
ν_5	0.079
ν_{14}	0.073
λ_5	-0.242
λ_{14}	-0.052

English proficiency	
Param.	Regularized Effect
ν_2	0.107
ν_5	0.025
ν_{13}	0.053
ν_{14}	-0.447
ν_{16}	-0.004
ν_{18}	-0.013
ν_{40}	0.120
λ_{13}	0.060
λ_{14}	0.469
λ_{40}	0.065

Unif. DIF

Nonunif. DIF

Applied Example: Results

- Identifying impact effects on θ mean and variance parameters

FRLP	
Param.	Effect
μ	-0.008
σ^2	0.021 (exp = 1.021)

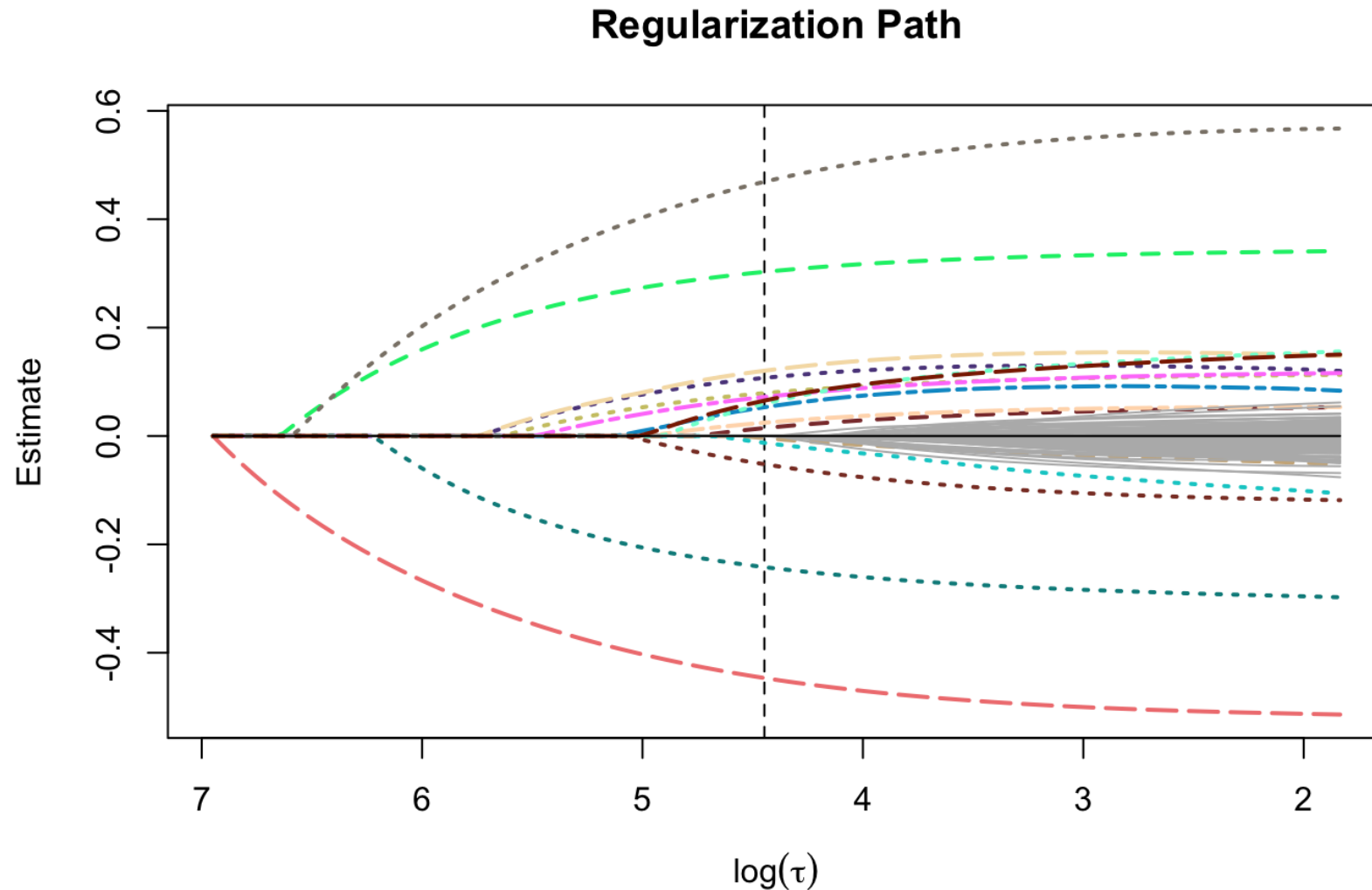
English proficiency	
Param.	Effect
μ	0.024
σ^2	0.022 (exp = 1.022)

$$\theta \sim N(0.024lang_i - 0.008FRLP_i, e^{0.022lang_i + 0.021FRLP_i})$$

- Expected θ slightly higher with language, slightly lower with FRLP
- θ variance slightly higher with both variables
- All effects look fairly small

Applied Example: Results

- Visualizing the regularization path



Further Steps and Questions

- Identify background effects of interest/concern
 - regDIF does not provide SE estimates
 - Need to determine minimum effect size of interest
 - Possibly estimate any non-zero effect and examine significance via ML
 - Can use cross-validation to improve robustness
- Sample size requirements for multiple background variables
 - Depends on the number of background variables and the complexity of the moderation equations for each parameter
 - The lack of cutoffs for sample size requirements
 - Can suffer from low power, especially for effects on discrimination parameters and variances

Further Steps and Questions (Cont.)

- Identify versus quantifying
 - If flagging DIF items is a primary goal, identifying non-zero paths may be sufficient
 - If making inferences is the goal, a Bayesian or ML approach might be a better option
 - Regularization is useful for variable selection but suboptimal for inference compared with a priori model specification
 - Need to align with goals of analysis
 - Strategize the sequence of multistage analyses to avoid overcomplexity or oversimplification of the model

References

- Bauer, D. J., Belzak, W. C. M., & Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(1), 43–55.
<https://doi.org/10.1080/10705511.2019.1642754>
- Belzak, W. C. M. (2023). The regDIF R package: Evaluating complex sources of measurement bias using regularized differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, 30(6), 974–984.
<https://doi.org/10.1080/10705511.2023.2170235>

MNLFA Code Walkthrough

5

5

MNLFA Code Walkthrough

Section Learning Objectives

Set up data for regularized MNLFA analysis

Perform analysis on prepared dataset

Interpret parameter estimates and information about estimation process

Determine next steps based on analytic goals

Applied Example: Data

- Simulated dataset
- $N = 10,000$ students
- 40 binary items (1 = correct) from a math ability test
- Binary measure of socio-economic status (0 = low, 1 = high)
- Gaussian IRT-derived score of language ability ($\mu_{\theta_l} = 0, \sigma_{\theta_l}^2 = 1$)


Applied Example: Data

- N = 10,000 students **ROWS**
- 40 binary items (1 = correct) from a math ability test **COLUMNS**
- Binary measure of socio-economic status (0 = low, 1 = high) **VECTOR**
- Gaussian IRT-derived score of language ability ($\mu_{\theta_l} = 0, \sigma_{\theta_l}^2 = 1$) **VECTOR**

Running Lasso Regularized MNLFA

With `items` the dataframe of responses, `ses` the vector of SES values, and `language` the vector of language values:

```
library(regDIF)
predictors = data.frame(ses = ses, language =
language)
fit.mnlfa = regDIF(items, predictors, num.tau = 100)
```



The number of different values of the penalty
term in lasso regularization
Lasso is the default for regDIF

Running Elasticnet Regularized MNLFA

The code to use a different regularization methods is simple, though it does require some familiarity with the available methods (Belzak, 2023):

```
library(regDIF)
predictors = data.frame(ses = ses, language =
language)
fit.mnlfa = regDIF(items, predictors, num.tau = 100,
alpha = 0.5)
```

Results

Once the code has run, we can review results in a few ways.

```
summary(fit.mnlfa)
```

Call:

```
regDIF(item.data = items, pred.data = predictors, num.tau = 100)
```

Optimal model (out of 100):

tau	bic
85.37939	507267.30570

Non-zero DIF effects:

item2.int.ses	item2.int.language	item3.int.ses	item5.int.ses	item5.int.language	item13.int.language
0.0146	0.1072	0.3028	0.0791	0.0245	0.0531
item14.int.ses	item14.int.language	item16.int.language	item18.int.language	item40.int.language	item5.slp.ses
0.0720	-0.4466	-0.0039	-0.0127	0.1204	-0.2420
item13.slp.language	item14.slp.ses	item14.slp.language	item40.slp.language		
0.0595	-0.0521	0.4691	0.0653		

Results

We can dig deeper on impact:

```
> mnlfa_coef <- coef(fit.mnlfa, tau = "tau.min")
```

```
> mnlfa_coef$impact
```

mean.ses	mean.language	var.ses	var.language
-0.0080	0.0236	0.0210	0.0222

Results

We can dig deeper on item parameters:

```
> mnlfa_coef <- coef(fit.mnlfa, tau = "tau.min")
> mnlfa_coef$base
```

item1.int.	item2.int.	item3.int.	item4.int.
-1.1742	-0.3899	0.0802	-0.6622

.....

item1.slp.	item2.slp.	item3.slp.	item4.slp.
0.3869	0.5663	0.4934	0.3791

Results

We can dig deeper on DIF:

```
> mnlfa_coef <- coef(fit.mnlfa, tau = "tau.min")
```

```
> mnlfa_coef$dif
```

item1.int.ses	item1.int.language	item2.int.ses	item2.int.language
0.0000	0.0000	0.0146	0.1072

Results

- We've focused on the most important parts of the output but there is a lot more included.

```
fit.mnlfA Large regDIF (15 elements, 22.5 MB)
$ tau_vec : num [1:100] 1042 1011 980 950 921 ...
$ aic : num [1:100] 508095 508071 508046 508023 508001 ...
$ bic : num [1:100] 508701 508684 508659 508636 508614 ...
$ impact : num [1:4, 1:100] 0.0129 0.0293 0.0116 0.043 0.0129 0.0294 0.0115 0.043 0.0129 ...
..- attr(*, "dimnames")=List of 2
.. ..$ : chr [1:4] "mean.ses" "mean.language" "var.ses" "var.language"
.. ..$ : NULL
$ base : num [1:80, 1:100] -1.1759 -0.3926 0.0749 -0.664 -0.0297 ...
..- attr(*, "dimnames")=List of 2
.. ..$ : chr [1:80] "item1.int." "item2.int." "item3.int." "item4.int." ...
.. ..$ : NULL
$ dif : num [1:160, 1:100] 0 0 0 0 0 0 0 0 0 0 ...
..- attr(*, "dimnames")=List of 2
.. ..$ : chr [1:160] "item1.int.ses" "item1.int.language" "item2.int.ses" "item2.int.language" ...
.. ..$ : NULL
$ eap :List of 2
..$ scores: num [1:10000, 1:100] -0.268 -1.169 -0.368 0.906 -0.467 ...
..$ sd : num [1:10000, 1:100] 0.565 0.574 0.559 0.564 0.568 ...
$ estimator_history:List of 100
..$ : num [1:245, 1:31] -0.977 0.761 0 0 0 ...
.. ..- attr(*, "dimnames")=List of 2
.. .. ..$ : chr [1:245] "c0_item1_int1" "a0_item1_" "c1_item1_cov1" "c1_item1_cov2" ...
.. .. ..$ : NULL
```

Next Steps

- What to do with this information, of course, depends on the context.
- Items with DIF may be flagged for removal, or not, depending on the setting.
- Results can be compared with those of a non-moderated 2PL analysis
 - Distributions of EAP scores may be of particular interest – helps assess influence of DIF on score distributions

Re-Estimating the Model

- One common next step is to re-estimate the model.
 - This is because regularization can identify salient vs. ignorable DIF parameters, but may also introduce bias into the estimates of non-zero paths.
 - Common to use the regularized model results to help specify an identified model in a software that performs maximum likelihood or Markov Chain Monte Carlo estimation, then yielding final parameter estimates.
 - Activity includes Mplus example
- Resources for estimation in other software:
 - Kolbe et al. (2024): *OpenMX* (ML, continuous indicators)
 - Enders et al. (2024): *Blimp* (MCMC, continuous or categorical indicators)
 - Bauer (2017): *Mplus* (ML, categorical shown but continuous also fine)

Conclusion

- This concludes our module.
- Code to replicate the example analysis, an activity expanding on it, and resources for additional reading are available online.
- Thank you!

References

- Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, 22(3), 507–526. <https://doi.org/10.1037/met0000077>
- Belzak, W. C. M. (2023). The regDIF R package: Evaluating complex sources of measurement bias using regularized differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, 30(6), 974–984. <https://doi.org/10.1080/10705511.2023.2170235>
- Enders, C. K., Vera, J. D., Keller, B. T., Lenartowicz, A., & Loo, S. K. (2024). Building a simpler moderated nonlinear factor analysis model with Markov Chain Monte Carlo estimation. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000712>
- Kolbe, L., Molenaar, D., Jak, S., & Jorgensen, T. D. (2024). Assessing measurement invariance with moderated nonlinear factor analysis using the R package OpenMx. *Psychological Methods*, 29(2), 388–406. <https://doi.org/10.1037/met0000501>