Item Response Theory models for Detection of Differential Item Functioning

Tolu Sajobi

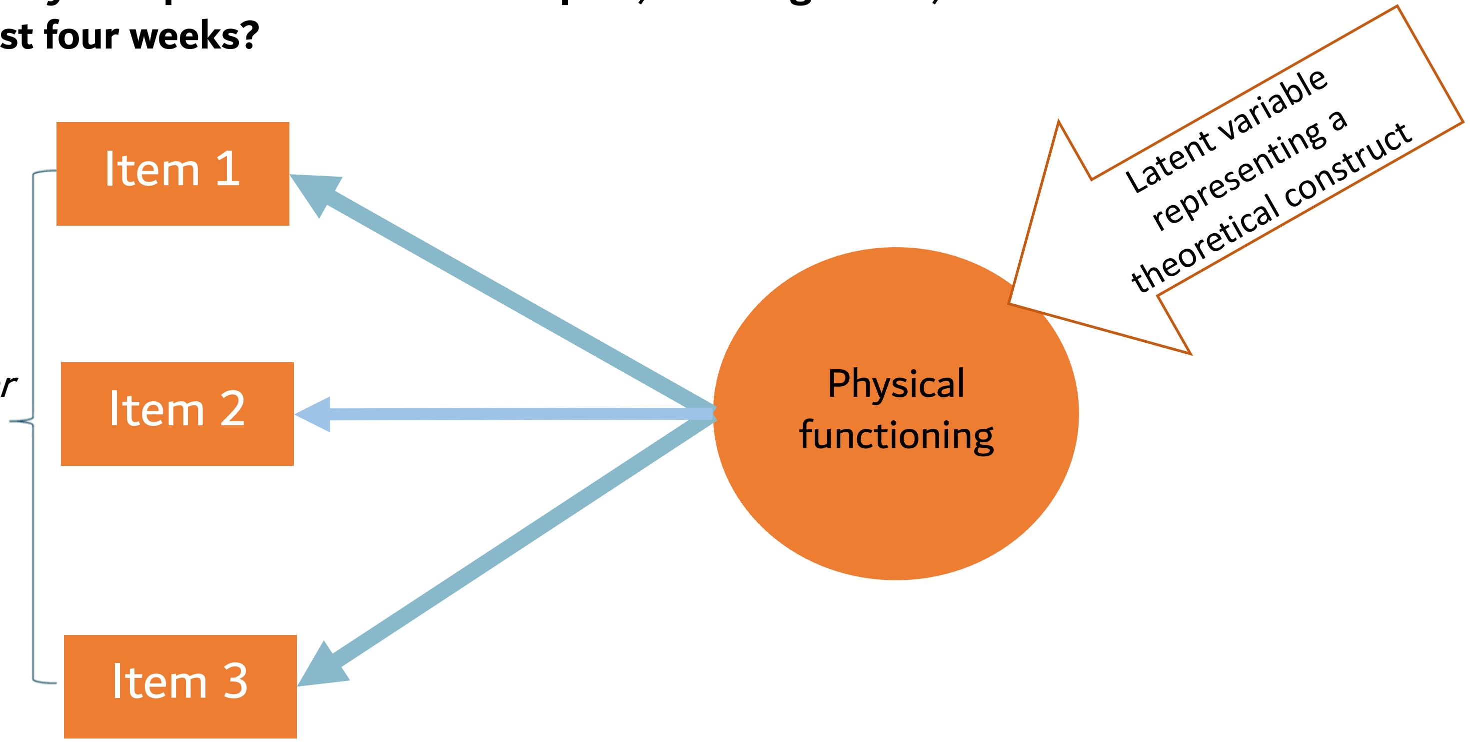
Item Response Theory Models

A class latent variable models that models the relationship between observed responses and the theoretical construct

Observations / measurement indicators /items

How much limitations have you experience due to chest pain, chest tightness, or angina over the the past four weeks?

- *Walking indoor on level ground*
- *Gardening, vacuuming, or carrying groceries*
- *Lifting or moving heavy objects (e.g., furnitures, children)*



Example for the measurement of physical limitation based on Seattle Angina Question (SAQ-7) physical limitations subscale

IRT Models for Binary Items

IRT model	Description
1-parameter logistic model / RASCH	• Most basic model with a single difficulty parameter (b) • Loadings/discrimination (a) are fixed at 1
2-parameter logistic model	• Includes a difficulty parameter (b) and a discrimination parameter (a)
3-parameter logistic model	• Adds a 'guessing' or change parameter (c) (i.e., probability of 'success' even at lowest level of ability >0)

IRT Models for Polytomous items

IRT model	Description
Graded response model (2-parameter IRT model)	• Relationship between the items and the factor are defined by a logistic proportional odds model
Partial credit model (RASCH)	• Specification is the same was the GRM except that the loadings (discrimination) are set to be equivalent of all items
Generalized partial credit model (RASCH)	• Based on the PCM but allows for discrimination parameters to vary across items
Rating Scale Model	• Same rating scale category structure across items • Same Distance between categories on the logit scale • Same number of categories across items • Thresholds can be disordered

Graded Response Model

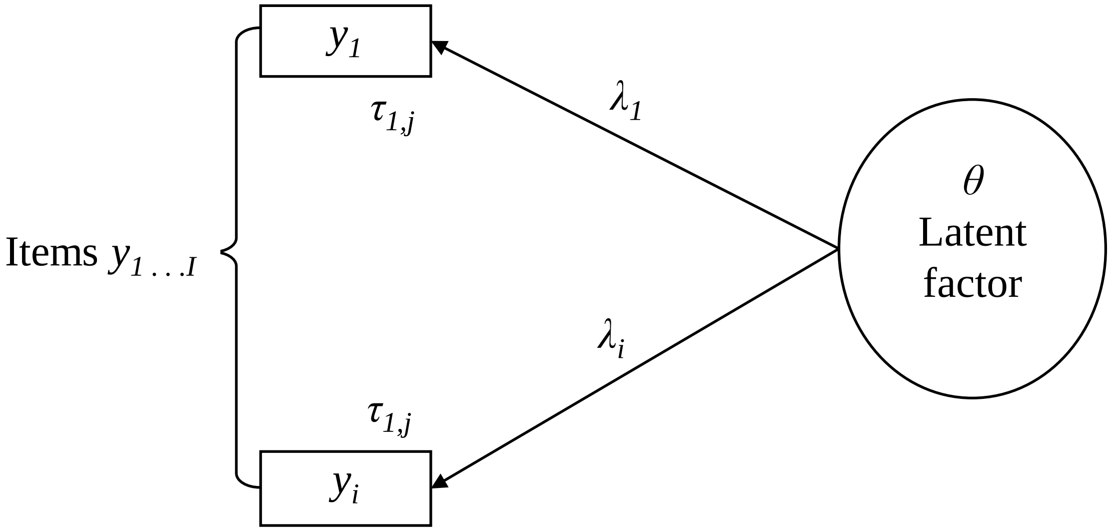
- Graded Response Model
  - The cumulative probability (P_{ij}) that the response to item I is at or above category j is

$$P_{ij}\left(Y \geq j \mid \theta\right) = \frac{\exp(-\tau_{ij} + \lambda_i \theta)}{1 + \exp(-\tau_{ij} + \lambda_i \theta)}$$

where

λ = factor loadings for items y_i, $i = 1,\ldots, I$.

τ = thresholds for $j - 1$ response categories per item



Graded Response Model for Polytomous Items

- The GRM can be parametrized as

$$P_{ij}\left(Y \geq j \mid \theta\right) = \frac{\exp(\alpha_i\left(\theta - \beta_{ij}\right))}{1 + \exp(\alpha_i\left(\theta - \beta_{ij}\right))}$$

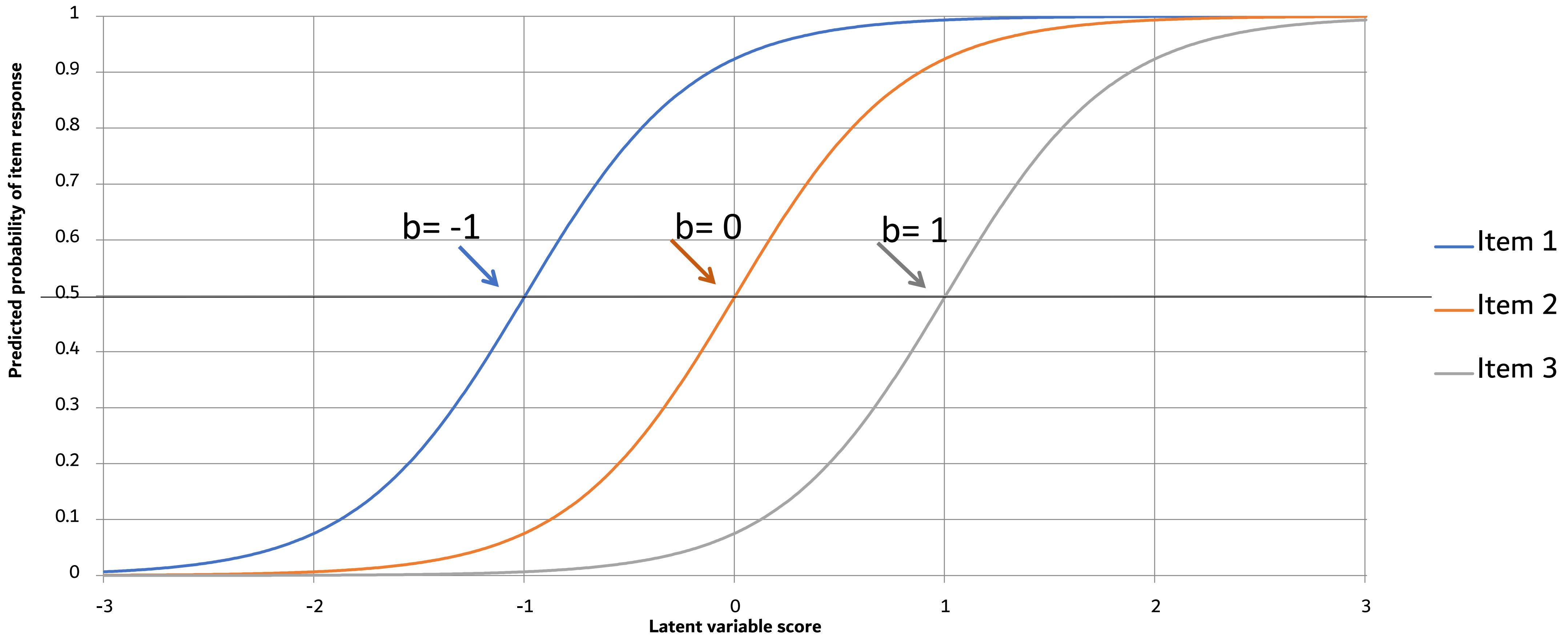
where

α = the discrimination parameter for item *i*
β = the difficulty parameter for the response categories less one

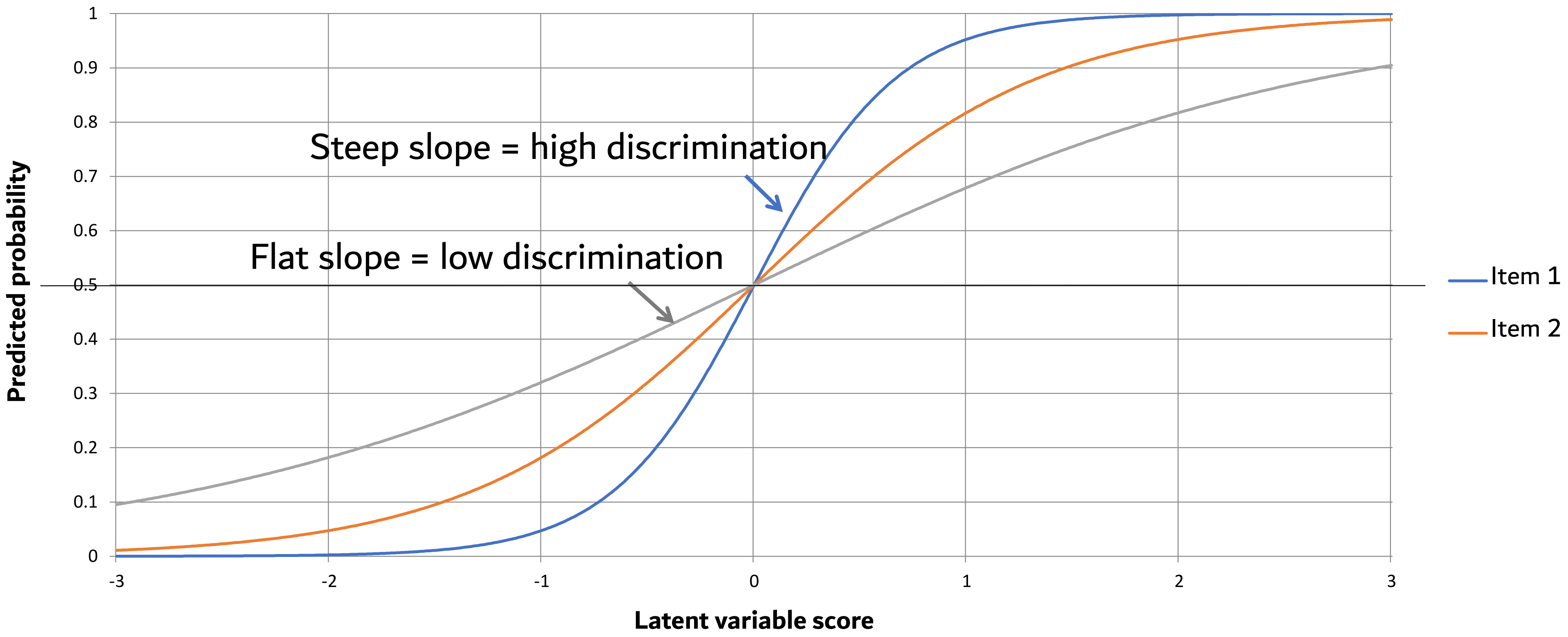
$$\beta_{ij} = \frac{\tau_{ij}}{\lambda_i} \qquad \alpha_i = \lambda_i$$

if θ is normally distributed with a mean of zero and variance of one, none of the thresholds or factor loadings are constrained, and a logistic link function with maximum likelihood estimation is used.

b-Parameter (difficulty/threshold)



a-Parameter (discrimination)



Assumptions of IRT Models

- Unidimensionality
- Local Independence
- Monotonicity
- Item Invariance

1. Unidimensionality

- Unidimensionality assumes that a set of items on a scale measure just one thing in common.

- Assessment of unidimensionality of PROMs could done via
  - exploratory factor analysis that identifies only one principal factor.
  - assessment of IRT model fit
  - Parallel analysis

2. Local Independence

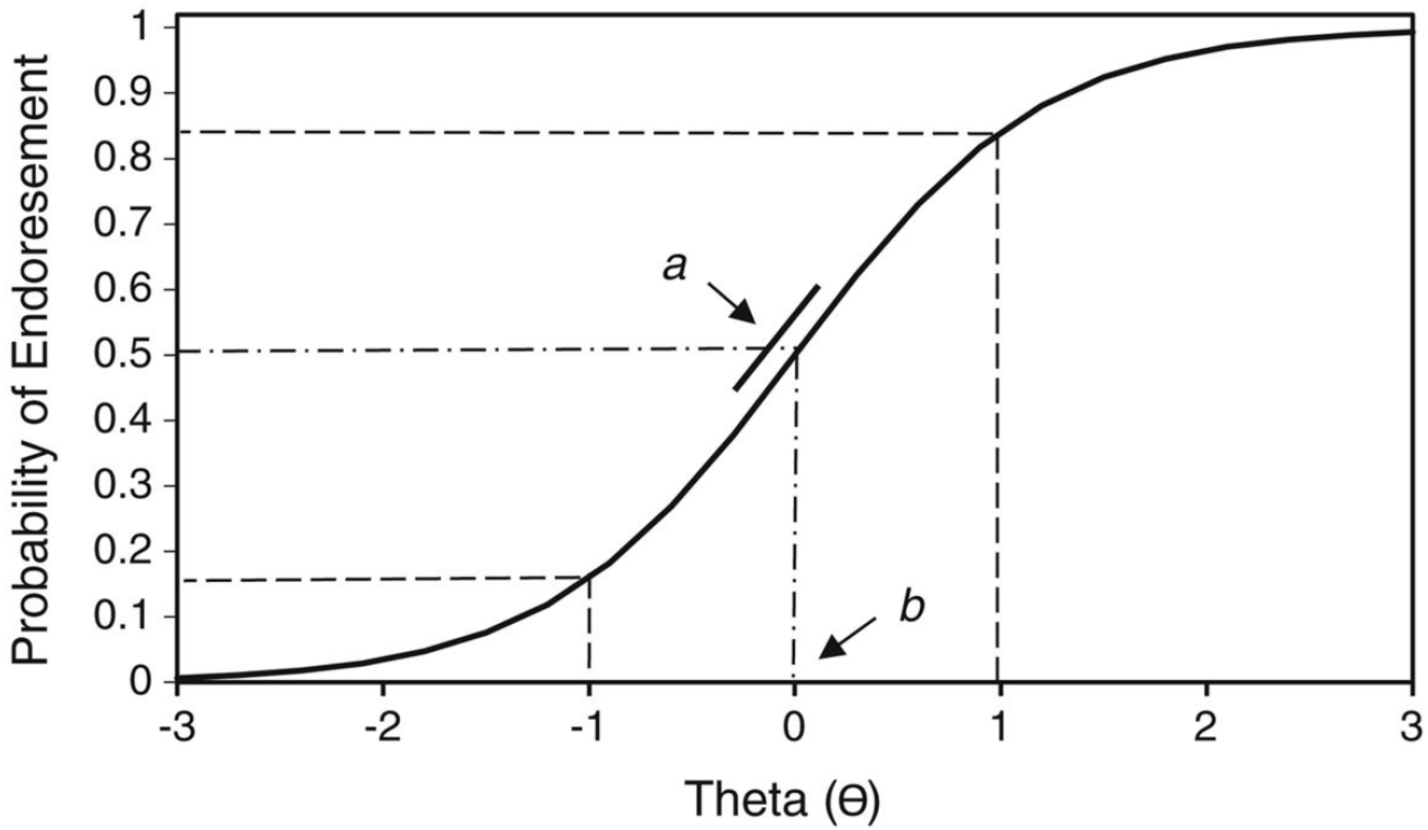
- Each and every item on a PRO measure is statistically independent of responses to all other items on the measure, conditional upon the latent trait.

- That is, conditional on the latent trait, responses on any pair of items are uncorrelated

- Violations of this assumptions can be tested by examining
  - Large magnitude of discrimination parameter (a > 4) for an item relative to other items
  - Residual covariance matrices to identify items with excessive covariation

3. Monotonicity

Item Characteristic Curve



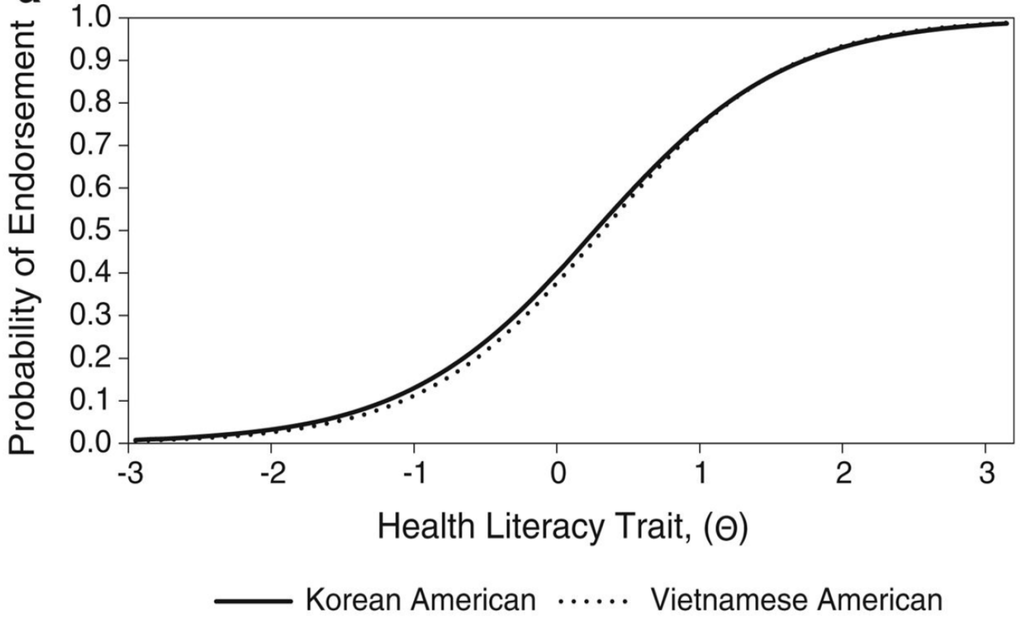
- The probability of endorsing an item increases as θ increases

Nguen TH, Han H, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. Patient 2014; 7(1): 23-35.

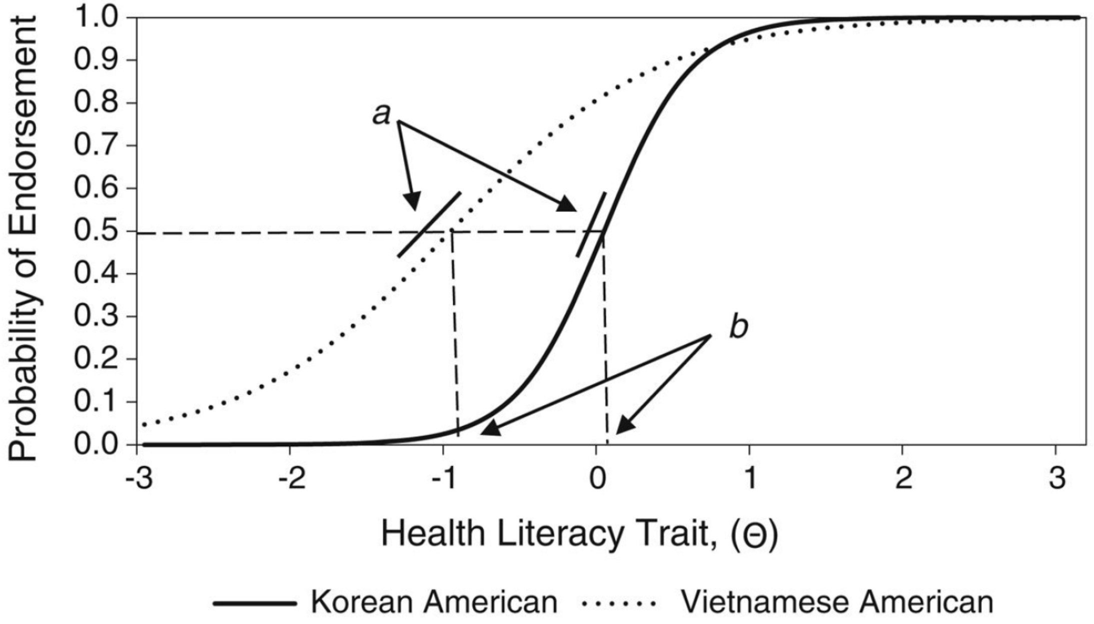
4. Item invariance

- The IRT model parameters are invariant across different populations
  - Violation of this assumption is known as **differential item functioning (DIF)**
  - For an item without DIF, the item characteristic curve is the same regardless population subgroup

Item Invariance



Item non-invariance



Nguen TH, Han H, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. Patient 2014; 7(1): 23-35.

Fit Indices for IRT Models

- Likelihood Ratio Statistic

$$G^2 = 2\sum_{k=1}^{K} O_i(k)\log\frac{O_i(k)}{E_i(k)} \sim \chi^2$$

- Standardized root mean square residual (SRMSR, Maydeu-Olivares & Joe, 2014)
  - SRMSR $\leq$ 0.05 indicates an acceptable fit
- Root Mean Square Error of Approximation (RMSEA)
  - RMSEA < 0.08 indicates acceptable fit

Fit Indices for IRT Models

- Information Theoretic Measures
  - Akaike Information Criterion
  - Bayesian Information Criterion

- These provide information about model fit relative to the number of model parameters

- Lower AIC and BIC values are indicative of better fit

Fit Indices for IRT Models

- Chi square test

$$\chi^2 = \sum_{k=1}^{K} \frac{[O_{ik} - E_{ik}]^2}{E_{ik}}$$

where K is the number of response categories for an item, O_{ik} is the observed frequency of endorsing option k, and E_{ik} is the expected frequency of option k under the IRT model.

- Limitations:
  - chi-square statistic is sensitive to sample size
  - test at the individual item level is insensitive to certain types of model misfits (Van den Wollenberg, 1982)

- Alternative tests
  - $S - \chi^2$ test

Mean Square Fit Indices for Rasch Models

- Response Residuals

$$Z_{ij} = \frac{Y_{ij} - E(Y_{ij})}{\sqrt{Var(Y_{ij}}}$$

$$Infit_i = \frac{\sum_{i=1}^{n} w_{ij} Z_{ij}^2}{\sum_{i=1}^{n} w_{ij}}$$

$$Outfit_i = \sum_{i=1}^{n} \frac{z_{ij}^2}{n}$$

- Rule-of-thumb for acceptable model fit:
  - 2.0 ≤Infit/outfit values ≤ 2.0. (Linacre, 2017)
  - $1 - \frac{6}{\sqrt{n}} \leq$ Outfit values $\leq 1 + \frac{6}{\sqrt{n}}$ (Smith et al, 1998)
  - $1 + \frac{2}{\sqrt{n}} \leq$ Outfit values $\leq 1 + \frac{2}{\sqrt{n}}$ (Smith et al, 1998)

Differential Item Functioning (DIF)

- Differential item functioning (DIF)
  - Probability of responses differs across respondents at the same level of the latent variable
- For example, male and female patients with the same level of health status rate their chest pain differently due to their interpretation of chest pain
- “Measurement Invariance” = “no Differential Item Functioning”

Differential Item Functioning

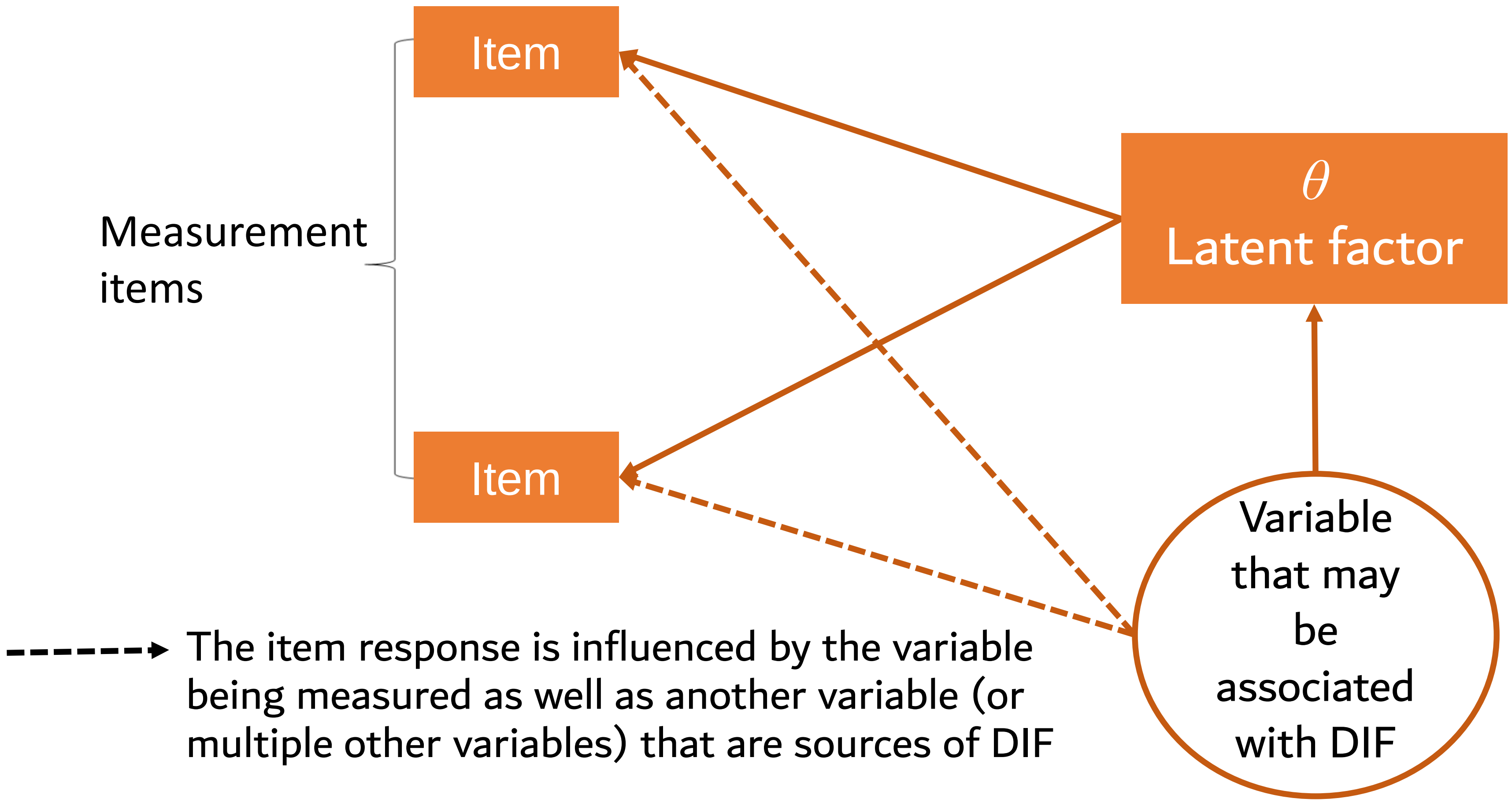
Assumption: Measurement instrument score reflects construct of interest

But - people may interpret and respond to questions in systematically unique ways because of:

- Demographics
- Culture
- Life circumstances and experiences
- Health experiences
- Personality

Reality: Measurement instrument score reflects construct + something else

Differential Item Functioning



Why DIF?

- Rule out measurement artifacts as an explanation for score differences
- Evaluate comparability of translated/adapted measurement instruments
- Support fairness and equality in measurement
- Evaluate the comparability of PROMs scores across groups in epidemiological and randomized clinical trials
- Understand item response processes

IRT: Likelihood Ratio Test for DIF Detection

- Identify the grouping variable on which DIF is to be tested

- Compare the reduced and full model with using likelihood ratio test
  - M_1: Discrimination (factor loadings) and difficulty (threshold) parameters are constrained to be equal across the grouping variable

  - M_2: Parameters are assumed to vary across the grouping variables

$$G = -2\left(LL_{M_1} - LL_{M_2}\right) \sim \chi^2_{df}$$

where

p = Degrees of freedom which is the number of unconstrained parameters

LL_{M_1} = loglikelihood of fully constrained models

LL_{M_2} = loglikelihood of partially constrained models

Teresi, J. A.,et al. Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF):Applications (with illustrations) to measures of physical functioning ability and general distress. Quality of Life Research 2007; 16:43-68.

Practical Steps for Implementing DIF using IRT

1. Evaluation of IRT model assumptions and fit
2. Specify the grouping variable of interest
3. Identification of anchor items
4. Purification of the anchor set.
5. Evaluate non-anchor items for DIF
6. Evaluate the magnitude of DIF on each item

Limitations of the Multigroup IRT for DIF Detection

- This methodology require a priori specification of the variables associated with DIF.
- Evaluation of DIF using multigroup IRT requires the variable of interest to be categorical.
  - The determination of the optimal number of threshold for categorizing a continuous variable can be subjective
  - Loss of information and statistical power when continuous variables associated with DIF are are categorized
- The use of multigroup IRT for DIF detection can be prohibitive when there are multiple variables associated with DIF.

Alternative DIF Detection methods

- Unsupervised latent variables are an alternative class of methods for test DIF
  - Allow for simultaneous evaluation of DIF on multiple variables
  - No a priori knowledge of potential variables that may be associated with DIF
  - Control of familywise Type I error

Worked Example

References & Resources

1. Teresi, J. A.,et al. Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF):Applications (with illustrations) to measures of physical functioning ability and general distress. Quality of Life Research 2007; 16:43-68.

2. de Ayala, R.J. (2009). The Theory and Practice of Item Response Theory. New York: The Guilford Press.

3. Embretson, S. E. & Reise, S. P. (2000). Item Response Theory for Psychologists. New York: L. Erlbaum Associates.

4. Hambleton, R. K. and Jones, R. W. (1993), An NCME Instructional Module on: Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. Educational Measurement: Issues and Practice, 12: 38–47. doi: 10.1111/j.1745-3992.1993.tb00543.x

5. Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. Med Care, 38(9 Suppl), II28-II42.Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. Med Care, 38(9 Suppl), II28-II42.

6. Penfield, R. D. (2014). An NCME Instructional Module on Polytomous Item Response Theory Models. Educational Measurement: Issues and Practice, 33(1), 36-48. doi:doi:10.1111/emip.12023

7. Reise, S. P. (2014). Item Response Theory Theory. In L. Cautin & S. O. Lilienfeld (Eds.), The Encyclopedia of Clinical Psychology: John Wiley & Sons.