An Introduction to Item Response Theory Analysis Using 2-parameter Logistic Model

Dr. Wan Nor Arifin

Biostatistics and Research Methodology Unit
Universiti Sains Malaysia
wnarifin@usm.my / wnarifin.github.io

Last update: Sep 4, 2025

Outlines

- Introduction
- Item Analysis
- Item Response Theory
- 2-parameter Logistic IRT
- Practical in R

Learning outcomes

- Understand the basic concepts in item response theory (IRT) analysis
- Perform 2-PL IRT analysis for dichotomous items

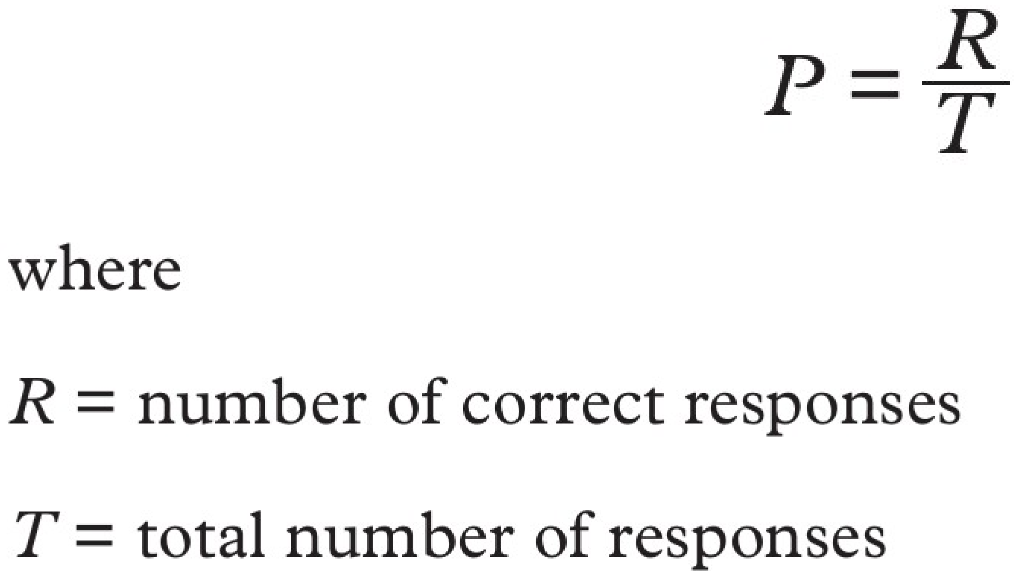
Introduction

What is Item Analysis (IA)

- Descriptive statistics
- Assess two components of test items:
  - Difficulty (P)
  - Discrimination (D)

What is Item Analysis (IA)

- Difficulty, P:



$$P = \frac{R}{T}$$

where

R = number of correct responses

T = total number of responses

What is Item Analysis (IA)

- Discrimination, D:

$$D = P_U - P_L$$

where

$$P_U = \frac{R_U}{T_U}$$

R_U = number of correct responses in the upper group (top 27% performers)

T_U = total number of responses in the upper group

$$P_L = \frac{R_L}{T_L}$$

R_L = number of correct responses in the lower group (bottom 27% performers)

T_L = total number of responses in the lower group

Practical

- Let’s calculate all these in Excel
- mtf.csv (Arifin & Yusoff, 2017)

What is Item Response Theory (IRT)

- Lord (1952) and Birnbaum (1968) → Foundation of IRT
- Model responses to items as interaction between item characteristics/parameters and a person's latent ability Meijer & Tendeiro (2018)
- Basis - Item characteristic curve (ICC)
  - "a logistic function that models the relationship between a person's response to an item and his/her level on the construct measured by the scale" Edelen & Reeve (2007)

What is Item Response Theory (IRT)

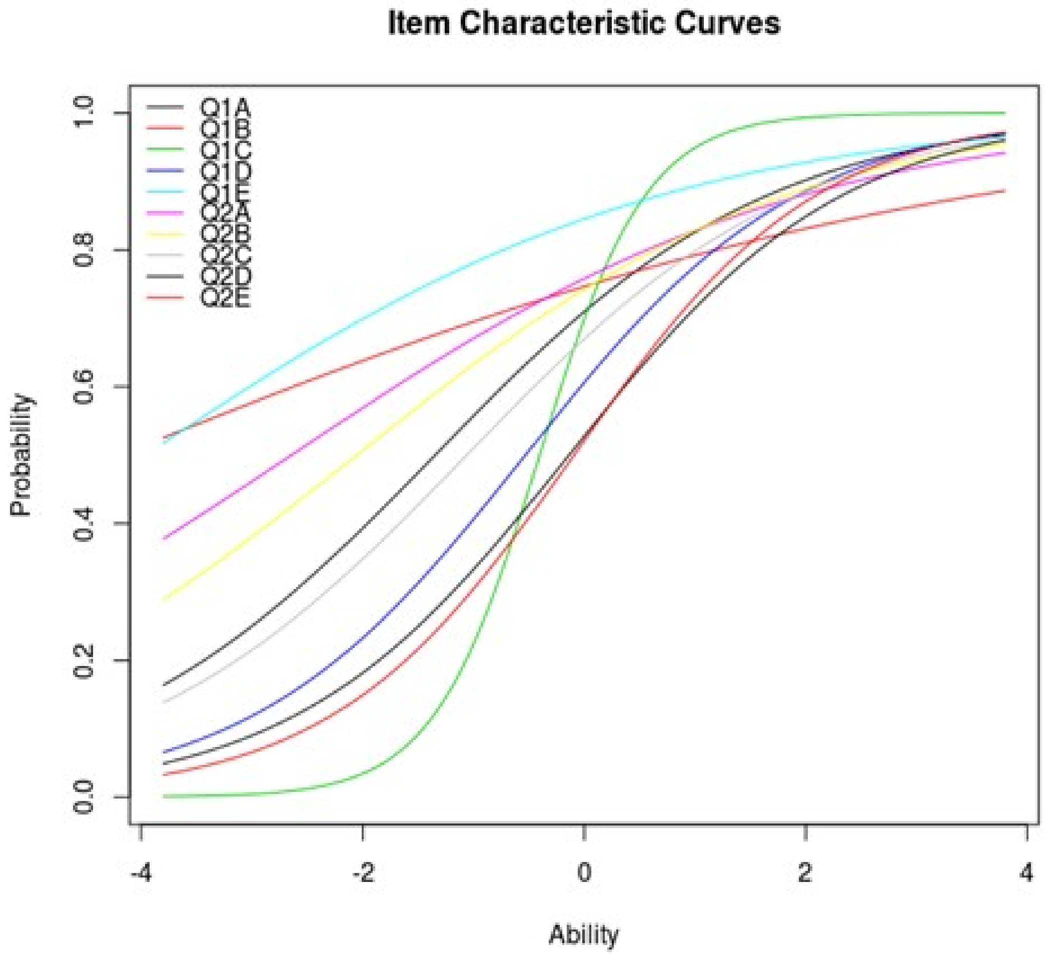


*Figure 1 - Arifin & Yusoff (2017)

What is Item Response Theory (IRT)

Models for dichotomous items by number of parameters:

1 parameter logistic (1PL)

- Difficulty (b)

2 parameter logistic (2PL) - common used Edelen & Reeve (2007)

- Difficulty (b), Discrimination (a)

3 parameter logistic (3PL)

- Difficulty (b), Discrimination (a), Guessing (c)

What is Item Response Theory (IRT)

Terms:

- Person's latent ability (θ)
  - Underlying ability level/score Baker (2001)
  - Latent trait/construct Reeve & Masse (2004)
- Difficulty (b)
  - Location, threshold - Point on ICC at which 50% respondents' get the item right
- Discrimination (a)
  - Slope at b threshold point on ICC
- Guessing (c)
  - Respondents' probability of getting an item correct by chance
  - Usually for education items Edelen & Reeve (2007)

IRT Model Selection

Two strategies[Meijer & Tendeiro (2018)]

1. Best fitting model with the smallest number of parameters for the data
2. Choose IRT model, then delete items that don’t fit

What is 2PL Model

- Birnbaum's 2PL model:

$$P(X_j=1|\theta, A_j, B_j)=\frac{\exp(A_i[\theta-B_j])}{1+\exp(A_j[\theta-B_j])}$$
$$=p_j$$

where
X_j, item response
θ, person's ability
A_j, item discrimination parameter
B_j, item difficulty parameter

CTT vs IRT*

- CTT
  - Scale: Numerical, categorical (dichotomous, polytomous)
  - Scale properties are sample dependent, rely on 1st and 2nd statistical moments (means, variances)Reeve & Masse (2004)
- IRT
  - Scale: Categorical (dichotomous, polytomous)
  - Scale properties are stable, not sample dependent, rely on higher order moments (e.g. threshold, slope parameters) → psychometrically invariantReeve & Masse (2004)

*For comprehensive comparisons, refer to Reeve & Masse (2004)

Categories of IRT Analysis Activities

Analysis Categories

Three categories of IRT analysis activities:

- Calibration
- Model-data fit
- Other validity evidence

Calibration

Three categories of analysis activities:

- Calibration
- Model-data fit
- Other validity evidence

Fit IRT model to estimate:

- Each Item **Difficulty, Discrimination**

Range:

Difficulty

-ve → zero → +ve

Easier → Middle → Difficult

Discrimination – 0.8 to 2.5 (Good)$^{\text{de Ayala}}$ (2009)

Model-data fit

Three categories of analysis activities:

- Calibration
- Model-data fit
- Other validity evidence

Before calibration:
Dimensionality assessment – unidimensionality (one dimension / trait)

- Factor analysis for categorical data
- EFA on tetrachoric correlations
- CFA using estimation methods that handle categorical data

After calibration:
Item & Person Fits
Model fit
Unidimensionality
Reliability

- Empirical reliability
- Item, test Information

Graphical assessment

Model-data fit

Three categories of Rasch analysis activities:

- Calibration
- Model-data fit
- Other validity evidence

- Parallel test
- Ratio of 1st:2nd eigenvalues > 3 Morizot et al (2007)

Before calibration:
Dimensionality assessment – unidimensionality (one dimension / trait)

- Factor analysis for categorical data
- EFA on tetrachoric correlations
- CFA using estimation methods that handle categorical data

After calibration:
Item & Person Fits
Model fit
Unidimensionality
Reliability
Empirical reliability
Item, test Information
Graphical assessment

Model-data fit

Three categories of Rasch analysis activities:

- Calibration
- Model-data fit
- Other validity evidence

Before calibration:
Dimensionality assessment – unidimensionality (one dimension / trait)

- Factor analysis for categorical data
- EFA on tetrachoric correlations
- CFA using estimation methods that handle categorical data

After calibration:
Item & Person Fits
Model fit
Unidimensionality

- Reliability
- Empirical reliability

Item, test Information
Graphical assessment

- Item characteristic curve (ICC)
- Item and test information curve

Other validity evidence

Three categories of Rasch analysis activities:

- Calibration
- Model-data fit
- Other validity evidence

- Invariance of item parameters
- Differential item functioning (DIF)
- Other typical construct validity evidence

Other validity evidence

Three categories of Rasch analysis activities:

- Calibration
- Model-data fit
- Other validity evidence

- Invariance of item parameters
- Differential item functioning (DIF)
- Other typical construct validity evidence

- Split sample into two-halves randomly
- Fit IRT model
- Correlate between two sample estimates

Other validity evidence

Three categories of Rasch analysis activities:

- Calibration
- Model-data fit
- Other validity evidence

- Invariance of item parameters
- Differential item functioning (DIF)
- Other typical construct validity evidence

- Whether performance on any of the items differs for certain groups (e.g. male vs female)
- Probability of correctly responding to an item should be the same for males and females

Other validity evidence

Three categories of Rasch analysis activities:

- Calibration
- Model-data fit
- Other validity evidence

- Invariance of item parameters
- Differential item functioning (DIF)
- Other typical construct validity evidence

Comparison vs known criteria, other instruments/variables

2PL IRT Analysis in R

Practical

- Let’s obtain all these in R
- mtf.csv (Arifin & Yusoff, 2017)
- practical_irt_2pl.html (tutorial in R)

References

- Arifin, W. N., & Yusoff, M. S. B. (2017). Item response theory for medical educationists. *Education in Medicine Journal*, 9(3), 69–81.
- Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. The Guilford Press.
- Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16, 5–18.
- Mair, P. (2018). *Modern psychometrics with R*. Springer.
- Meijer, R. R., & Tendeiro, J. N. (2018). Unidimensional item response theory. In P. Irwing, T. Booth, & D. J. Hugh (Eds.), *The Wiley handbook of psychometric testing : A multidisciplinary reference on survey, scale and test development* (pp. 413-433). Wiley.
- Morizot, J., Ainsworth, A. T., & Reise, S. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), Handbook of research methods in personality psychology (pp. 407–423). New York: Guildford.
- Reeve, B. B., & Ma ˆsse, L. C. (2004). Item response theory modeling for questionnaire evaluation. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Sinter (Eds.), *Methods for testing and evaluation survey questionnaires* (pp. 247–273). Hobeken, NJ: Wiley.