Item response theory (IRT)

- Limitations of classical measurement models
- Defining IRT
- Assumptions of IRT
- Popular dichotomous models of IRT
- Requirements of IRT
- Popular polytomous models of IRT
Limitations of classical measurement models

- Test-dependent ability scores and group-dependent item statistics
- Reliability and standard error of measurement
- Test oriented than item oriented
- Less-than-ideal solutions to testing problems
What is IRT?

- The relationship between the ability measured by an instrument and an item response
- Categorical item responses
- Polytomous item responses
- Explanation of an examinee’s performance on a test item by a set of factors
- Item characteristic curve or function to show the relationship between the examinee’s performance and the set of factors
Assumptions of IRT

- Unidimensionality
- Local independence
- Correct model specifications
Unidimensionality

- The presence of a dominant component or factor

- Psychometric versus psychological unidimensionality

- Eigenvalues to test unidimensionality
An example of a scree plot to show unidimensionality
Local independence or conditional independence

- Complete latent space
- More than one ability, presence of clues, helpful information of the item
- Unidimensionality and local independence
Correct model specifications (fit)

- Model-to-data fit
- Items and persons fit
- Residuals to check for fit
Examples of fit and misfit
Models of IRT

- Models for dichotomous items
  1. One-parameter logistic model
  2. Two-parameter logistic model
  3. Three-parameter logistic model

- Models for polytomous items
  1. Graded response model
  2. Multi-faceteded Rasch model
Item difficulty, item discrimination, and guessing

- Item difficulty in classical and IRT models (P and b)
  Difficult items have lower difficulty indices in CM, but difficult items have higher difficulty indices in IRT models.

- Item discrimination in classical and IRT models (Point-biserial correlation and a)
  In both models, higher values indicate higher discrimination.
Models for dichotomous items

- Three-parameter logistic model
  1. Birnbaum (1968)
  2. The presence of all three parameters
  3. The discrimination parameter, a, shows the steepness of the curves. (0,2)
  4. The difficulty parameter, b, shows how difficult a particular item is. (-0.2, +2.0)
  5. Guessing or pseudo-chance parameter provides a nonzero asymptote for item characteristic curve.
Two items with differing discrimination

![Graph showing probability vs. \( \theta \) for two items with differing discrimination.](image)
Two items with differing difficulty

![Graph showing the probability of correctly answering items 1 and 2 as a function of \( \theta \).]
Items with differing lower asymptotes

![Graph showing probability against theta for two items](image)
Three items with the same discrimination but differing levels of difficulty.
Three items with the same difficulty but differing levels of discrimination
An item with perfect discrimination
Two-parameter logistic model (2PL)

- Lord (1952)
- The presence of only two parameters
- Item discrimination (a)
- Item difficulty (b)
- Guessing fixed to zero
An example of 2PL with different levels of a and b
Another example of 2PL
One-parameter logistic model (1PL)

- Rasch (1960/1980)
- The presence of only one parameter
- Difficulty parameter (b)
- Guessing fixed to zero
- All items to be equally discriminating
An example of 1PL for four items
Another example of 1PL
Requirements of dichotomous IRT models

- Data sample design—no normal distribution

- Data requirements
  1. 3PL—1000 examinees and 60 items
  2. 2PL—500 examinees and 30 items
  3. 1PL—100 examinees and 20 items
Models for polytomous data

- Graded response model
- Multi-facted Rasch model
Graded response model

- Samejima (1969)
- Applicable to ordered categories, using a scoring rubric
  1. A 5-point likert scale (strongly disagree, disagree, neutral, agree, strongly agree)
  2. A 6-point rating scale (very poor, poor, fair, good, very good, excellent)
- Appropriate for products (essays, portfolios, and presentations)

- The probability of getting a category increases as the ability increases

- $b$-parameters are called thresholds
<table>
<thead>
<tr>
<th>Essay number:</th>
<th>Rater's name:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Very poor</td>
</tr>
<tr>
<td>1. Substance</td>
<td>1 □</td>
</tr>
<tr>
<td>2. Thesis development</td>
<td>1 □</td>
</tr>
<tr>
<td>3. Topic relevance</td>
<td>1 □</td>
</tr>
<tr>
<td>4. Introduction</td>
<td>1 □</td>
</tr>
<tr>
<td>5. Coherent support</td>
<td>1 □</td>
</tr>
<tr>
<td>6. Conclusion</td>
<td>1 □</td>
</tr>
<tr>
<td>7. Logical sequencing</td>
<td>1 □</td>
</tr>
<tr>
<td>8. Range</td>
<td>1 □</td>
</tr>
<tr>
<td>9. Word choice</td>
<td>1 □</td>
</tr>
<tr>
<td>10. Word form</td>
<td>1 □</td>
</tr>
<tr>
<td>11. Sentence variety</td>
<td>1 □</td>
</tr>
<tr>
<td>12. Overall grammar</td>
<td>1 □</td>
</tr>
<tr>
<td>13. Spelling</td>
<td>1 □</td>
</tr>
<tr>
<td>14. Essay format</td>
<td>1 □</td>
</tr>
<tr>
<td>15. Punctuation/capitalization/handwriting</td>
<td>1 □</td>
</tr>
</tbody>
</table>
An example of a graded response model
Family of Rasch models

- Basic Rasch model or standard dichotomous model
- Andrich’s (1978) rating scale model
- Master’s (1982) partial credit model