* Item Response Theory
...in 1 hour

David Stillwell, ds617@cam.ac.uk

www.psychometrics.cam.ac.uk
Understanding the Latent Variable
* Personality scales?
  * Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism
* Intelligence scales?
  * Numerical Reasoning, Digit Span
* Ability scales?
  * Find egs. Of ability

*What is a construct?
* Measures and items are created in order to measure/tap a construct

* A construct is an underlying phenomenon within a scale - referred to as the **latent variable** (LV)

  * *Latent*: not directly observable
  * *Variables*: aspects of it such as strength or magnitude, change
  * Magnitude of the LV measured by a scale at the time and place of measurement is the **true score**

*What is a construct?*
* LV is regarded as a *cause* of the item score
  * i.e. Strength of LV (its true score) causes items to take on a certain score.
* Cannot directly assess the true score
  * Therefore look at correlations between the items measuring the same construct
  * Invoke the LV as cause of these correlations
  * Infer how strongly each item correlated with LV

* **Latent variable as the cause of item values**
Observed Test Score = True Score + random error

* Item difficulty and discrimination
* Reliability

Limitations:
* Single reliability value for the entire test and all participants
* Scores are item dependent
* Item stats are sample dependent
* Bias towards average difficulty in test construction

Classical Test Theory
*Introduction to IRT*

Some materials and examples come from the ESRC RDI in Applied Psychometrics run by:

Anna Brown (University of Cambridge)
Jan Böhnke (University of Trier)
Tim Croudace (University of Cambridge)
Ratio of correct responses to items on different level of total score
Item Response Function

Binary items

Parameters:
• Difficulty
• Discrimination
• Guessing
• Inattention

Models:
• 1 Parameter
• 2 Parameter
• 3 Parameter
• 4 Parameter
• unfolding

Probability of getting item right

Measured concept (theta)
One-Parameter Logistic Model/Rasch Model (1PL)

7 items of varying difficulty (b)

*One-Parameter Logistic Model/Rasch Model (1PL)*
Two-Parameter Logistic Model (2PL)

5 items of varying difficulty (b) and discrimination (a)
Three-Parameter Model (3PL)

One item showing the guessing parameter (c)

*Item 1: b=0.0, a=1.0, c=0.2*
*Fisher Information Function*
*(Fisher) Test Information Function*

Three items
Error of measurement inversely related to information

Standard error (SE) is an estimate of measurement precision at a given theta.
Test:
1. Normal distribution
2. q1 - Correct
3. q2 - Correct
4. q3 - Incorrect
**Classical Test Theory vs. Item Response Theory**

<table>
<thead>
<tr>
<th></th>
<th>Classical</th>
<th>IRT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Modelling / Interpretation</strong></td>
<td>Total score</td>
<td>Individual items (questions)</td>
</tr>
<tr>
<td><strong>Accuracy / Information</strong></td>
<td>Same for all participants and scores</td>
<td>Estimated for each score / participant</td>
</tr>
<tr>
<td><strong>Adaptivity</strong></td>
<td>Virtually not possible</td>
<td>Possible</td>
</tr>
<tr>
<td><strong>Score</strong></td>
<td>Depends on the items</td>
<td>Item independent</td>
</tr>
<tr>
<td><strong>Item Parameters</strong></td>
<td>Sample dependent</td>
<td>Sample independent</td>
</tr>
<tr>
<td><strong>Preferred items</strong></td>
<td>Average difficulty</td>
<td>Any difficulty</td>
</tr>
</tbody>
</table>
Why use Item Response Theory?

* Reliability for each examinee / latent trait level
* Examinee / Item parameters on the same scale (so you can compare the outcomes of two different tests measuring the same latent trait)
* Examinee / Item parameters invariance (so if you test a latent trait on a subsection of the population you can get parameter estimates for the whole population)
* Adaptive testing
* Also, test development is: cheaper and faster! (no need to re-norm every time you change an item)
Introduction to CAT
Very brief
* Standard test is likely to contain questions that are too easy and/or too difficult

* Adaptively adjusting to the level of the test to this of participant:
  * Increases the accuracy
  * Saves time / money
  * Prevents frustration
Start the test:
1. Ask first question, e.g. of medium difficulty
2. Correct!
3. Score it
4. Select next item with a difficulty around the most likely score (or with the max information)
5. And so on…. Until the stopping rule is reached
* IRT model
* Item bank and calibration
* Starting point
* Item selection algorithm (CAT algorithm)
* Scoring-on-the-fly method
* Termination rules
* Item bank protection / overexposure
* Content Balancing

* Elements of CAT