Introduction to Item Response Theory and Computer adaptive testing

Chris Gibbons PhD.
Director of Health Assessment and Innovation
NIHR Fellow
psychometrics.cam.ac.uk
Today’s lecture

- Psychometric Theory
- Classical/Modern test theories
- Computer adaptive testing
Learning outcomes

- General understanding of classical test theory (CTT) and item response theory (IRT)
- Explain why IRT is superior to CTT
- Understand the concept of item difficulty
- Explain different IRT models and their assumptions
- Understand computer adaptive testing and how it works
- Develop and build a ‘hand made’ CAT in groups
Measurement theory

- Psychometrics
- Classical test theory
- Item response theory
Constructs
Psychometrics
What is a construct?

- A construct is an underlying phenomenon that a questionnaire measures - referred to as the **latent variable (LV)**
  - *Latent*: not directly observable
  - *Variables*: strength or magnitude can change
  - Magnitude of the LV measured by a scale at the time and place of measurement is the **true score**

- Measures and items are created in order to measure/tap a construct

- **Unidimensional** (usually)
What is a construct?

- Personality?
  - Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism

- Intelligence?
  - Numerical Reasoning, Digit Span

- Health?
  - Depression, fatigue, expectations, empowerment
Key concepts

- Reliability
- Validity
- Standardisation
- Calibration (e.g., clinical)
Reliability

• Reliability
  - Inter-rater reliability
  - Test-retest reliability
  - Internal consistency reliability

• Internal consistency
  - Cronbach’s Alpha (average inter-item correlation)
  - Marginal reliability (average conditional standard error)

• Test-retest
  - Correlation
  - t-test
  - Bland-Altman
Validity

- **Construct** (does it measure what it should)
- **Content** (does it cover a representative sample of the trait)
- **Criterion** (is it related to other similar constructs)
- **Concurrent** (is it related to the other similar measures)
- **Predictive** (can it predict scores on related measures)
- **Diagnostic** (can it be use for diagnosis)

Validity is assured when you develop items, before you conduct psychometric assessments
Standardisation

- Calculating norm referenced scores for the assessment
- e.g., IQ – 100 is always average, SD is 15
- Trait scores in personality psychology (0-100)
- Health assessment using PROMIS (0-100)
Calibration

• Understanding scale scores in relation to other constructs, educational levels, symptoms.

• Receiver operating characteristics (ROC curves)

• Co-calibration with other scales using IRT
Psychometric approaches

Item response theory

Classical test theory
Psychometric approaches

Probabilistic

Correlational
Psychometric approaches

Modern

Responses to Item 2

Probability of responding

Level of depression/Score on the scale

‘Classic’

Width

Length

0 10 20 30 40 50

0 20 40 60 80 100 120 140
Psychometric approaches

‘Hard’

‘Easy’
Psychometric approaches

‘Interesting’

‘Boring’
Classical Test Theory

Foundation:

- **Observed Test Score = True Score + random error**

- Assumptions/beliefs
  1. Item means are unaffected by error if there is a large number of respondents
  2. One item’s error is *not* correlated with another item’s error (stochastic/random error)
  3. Error terms are *not* correlated with the true score of the latent variable

- Scores are test and item dependent. Must administer all items
- Calibration is sample dependent – must have representative sample
Classical Measurement Assumptions

\[ X = T + e \]

- \( X \) = observed score
- \( T \) = true score
- \( e \) = error
Item response theory

- **Probabilistic** relationship between the questionnaire items and the people taking responding to them.

- The **more of an underlying** trait that the person has the **more likely they are to agree** to an item measuring that trait.

- Originally – the more able a person, the more likely they are to get an exam question right

- So the level of underlying trait is called **‘Ability’**

- The level of the trait that the item measured is called **‘Difficulty’**
**Probabilistic** approach to responding to items

Consider the following depression questionnaire:

<table>
<thead>
<tr>
<th></th>
<th>Question</th>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Some days I feel unhappy</td>
<td>Agree</td>
<td>Disagree</td>
</tr>
<tr>
<td>2</td>
<td>I don't enjoy things anymore</td>
<td>Agree</td>
<td>Disagree</td>
</tr>
<tr>
<td>3</td>
<td>I don't laugh anymore</td>
<td>Agree</td>
<td>Disagree</td>
</tr>
<tr>
<td>4</td>
<td>I want to die</td>
<td>Agree</td>
<td>Disagree</td>
</tr>
<tr>
<td>5</td>
<td>Sometimes I'm sad</td>
<td>Agree</td>
<td>Disagree</td>
</tr>
<tr>
<td>6</td>
<td>Life is difficult right now</td>
<td>Agree</td>
<td>Disagree</td>
</tr>
<tr>
<td>7</td>
<td>Things won't get better</td>
<td>Agree</td>
<td>Disagree</td>
</tr>
</tbody>
</table>
**Probabilistic approach to responding to items**

Consider the following depression questionnaire:

<table>
<thead>
<tr>
<th></th>
<th>Some days I feel unhappy</th>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>I don't enjoy things anymore</td>
<td>Agree</td>
<td>Disagree</td>
</tr>
<tr>
<td>3</td>
<td>I don't laugh anymore</td>
<td>Agree</td>
<td>Disagree</td>
</tr>
<tr>
<td>4</td>
<td>I want to die</td>
<td>Agree</td>
<td>Disagree</td>
</tr>
<tr>
<td>5</td>
<td>Sometimes I'm sad</td>
<td>Agree</td>
<td>Disagree</td>
</tr>
<tr>
<td>6</td>
<td>Life is difficult right now</td>
<td>Agree</td>
<td>Disagree</td>
</tr>
<tr>
<td>7</td>
<td>Things won't get better</td>
<td>Agree</td>
<td>Disagree</td>
</tr>
</tbody>
</table>

We can score Agree = 1 and Disagree = 0
Probabilistic approach to responding to items

<table>
<thead>
<tr>
<th>Person</th>
<th>Level of depression</th>
<th>Agree?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>14</td>
<td>1</td>
</tr>
</tbody>
</table>
Probabilistic approach to responding to items

<table>
<thead>
<tr>
<th>2</th>
<th>I don't enjoy things anymore</th>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
</table>

![Graph showing responses to Item 2](image)

<table>
<thead>
<tr>
<th>Person</th>
<th>Level of depression</th>
<th>Agree?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>14</td>
<td>1</td>
</tr>
</tbody>
</table>
Plotting item difficulty and person ability

<table>
<thead>
<tr>
<th>2</th>
<th>I don't enjoy things anymore</th>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Person</th>
<th>Level of depression</th>
<th>Agree?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>14</td>
<td>1</td>
</tr>
</tbody>
</table>
Plotting item **difficulty** and person **ability**

<table>
<thead>
<tr>
<th>2</th>
<th>I don't enjoy things anymore</th>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
</table>

Responses to Item 2

50% chance of Agreeing at Depression = 10.3

<table>
<thead>
<tr>
<th>Person</th>
<th>Level of depression</th>
<th>Agree?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>14</td>
<td>1</td>
</tr>
</tbody>
</table>
Plotting item difficulty and person ability

I don't enjoy things anymore

<table>
<thead>
<tr>
<th>Person</th>
<th>Level of depression</th>
<th>Agree?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>14</td>
<td>1</td>
</tr>
</tbody>
</table>

50% chance of Agreeing at Depression = 10.3
Level of depression
Fatigue
IQ
Anxiety
Neuroticism
Quality of life
Health Behaviour
Religiousness
Theta (θ)
Theta ($\theta$)

A common metric on which to talk about

- ‘Difficulty’ of items
- ‘Ability’ of persons
<table>
<thead>
<tr>
<th>Modelling / Interpretation</th>
<th>Classical</th>
<th>IRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total score</td>
<td></td>
<td>Individual items (questions)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accuracy / Information</th>
<th>Classical</th>
<th>IRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same for all participants and scores</td>
<td></td>
<td>Estimated for each score / participant</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Adaptivity</th>
<th>Classical</th>
<th>IRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virtually not possible</td>
<td></td>
<td>Possible</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Score</th>
<th>Classical</th>
<th>IRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depends on the items</td>
<td></td>
<td>Item independent</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item Parameters</th>
<th>Classical</th>
<th>IRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample dependent</td>
<td></td>
<td>Sample independent</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Preferred items</th>
<th>Classical</th>
<th>IRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average difficulty</td>
<td></td>
<td>Any difficulty</td>
</tr>
</tbody>
</table>
Item response function
A maths question

What is $4 \times 8$?

a.) 20
b.) 32
c.) 40
d.) 16
A maths question

What is 4 x 8?

Candidate One
A maths question

What is 4 x 8?

Candidate One

Probability close to 1
Person ability = 2.5
A maths question

What is 4 x 8?

Candidate Two
What is $4 \times 8$?

Candidate Two

Probability close to 0.1
Person ability = -2
A **maths** question

What is $4 \times 8$?

Candidate Three

Probability close to .5

Person ability = 0
A maths question

What is $4 \times 8$?

Candidate Three

Probability close to 0.5

Person ability = 0
A note about \( \theta \)

IRT models we have introduced so far are parametric (assume Gaussian/normal trait distribution)

Where \( \theta = 0 \) is the population mean then \( \pm 1 \theta = \pm 1 \) standard deviation
A note about $\theta$

IRT models we have introduced so far are parametric (assume Gaussian/normal trait distribution)

Where $\theta = 0$ is the population mean then $\pm 1 \theta = \pm 1$ standard deviation
Types of IRT model

• What models are there?
• How do they vary?

**Question**

What does theta ($\theta$) represent?

If an item has a high $\theta$ value, what does that mean in terms of difficulty?

If a person has a low $\theta$ value, what does that mean in terms of ability?
Item response function
Item ‘difficulty’ (b) parameter (1PL)

Theta where prob = .5

b = 0
Item ‘difficulty’ (b) parameter (1PL)
Item ‘difficulty’ (b) parameter (1PL)

Item difficulties
- Item 1 = -1
- Item 2 = -.5
- Item 3 = 0
- Item 4 = .5
- Item 5 = 1
Item ‘difficulty’ (b) parameter (1PL)
Item ‘discrimination’ (a) parameter

Gradient of slope
Where $p = 0.5$
$a = 1$
Item ‘discrimination’ (a) parameter
Item ‘discrimination’ (a) parameter
And item ‘difficulty’ (b) parameter
Guessing (c) parameter
Inattention (d) parameter
Unfolding (e) parameter
Different IRT Models

Non-parametric
- Mokken

1-parameter (1pl) models (only item difficulty changes)
- Rasch model (for dichotomous data)
- Partial Credit Model (for polytomous data)
- Rating Scale Model (for polytomous data)

2-parameter (3pl) models (item difficulty and discrimination changes)
- 2pl model
- Graded Response Model
- Generalized Partial Credit Model

3+ parameter (3pl) models (with a guessing/inattention/unfolding parameter)
- Three-parameter logistic model (etc..)

Multidimensional
- Compensatory
- Bi-factor
- ...

One parameter / Rasch model
Two parameter / IRT model
Polytomous characteristic curves
Item Response theory
Assumptions

- Nature of the item-category curve
- Scalability (monotonicity)
- Unidimensionality
- Local independence of items
- Responses caused solely by the underlying trait
IRT/Rasch analysis

• Assess item and model fit

Diagnose misfit and alter items to fit to model

• Dimensionality
• Category threshold ordering
• Local dependency
• Differential item functioning
• Reliability

• Export threshold values for computer adaptive testing!
Factor Structure

• Assess factor structure with CFA (if established scale) otherwise EFA works well

• EFA – Polychoric PCA with oblique rotation
Assessing fit

- Or we can investigate *why* the model is not fitting, using some known criteria that are liable to cause misfit
  - Dimensionality
  - Category threshold ordering
  - Local dependency
  - Differential item functioning

- We assess item and person fit to IRT model using a chi-square statistic (or INFIT/OUTFIT for the Rasch model)

- Non-significant chi square interaction (or INFIT/OUTFIT close to 1)

- Assess model fit using a chi-square or likelihood ratio test
How much do any difficulties in mobility bother you?

<table>
<thead>
<tr>
<th>Category</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not at all</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A little</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A moderate amount</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very much</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>An extreme amount</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Category **threshold ordering**

Probability vs. Theta $\theta$ graph
Category **threshold ordering**

<table>
<thead>
<tr>
<th>How much do any difficulties in mobility bother you?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not at all</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

![Graph showing probability distributions for different levels of mobility bother.](image)
How much do any difficulties in mobility bother you?

<table>
<thead>
<tr>
<th>Not at all</th>
<th>A little</th>
<th>A moderate amount</th>
<th>Very much</th>
<th>An extreme amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
### How much do any difficulties in mobility bother you?

<table>
<thead>
<tr>
<th>Category</th>
<th>1</th>
<th>2</th>
<th>2</th>
<th>3</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not at all</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A little</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A moderate amount</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Very much</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>An extreme amount</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Category threshold ordering**

![Graph showing probability distribution with categories and scores]
Local dependency

- Occurs when the response to an item is conditional on a response to another item
- Response to one item must not influence another
- *e.g.*, they are too similar
- Inflates reliability (reverse wording is bad practice!)
- Causes model misfit

How limited are you when –

- Walking more than a kilometer
- Walking more than half a kilometer
- Walking more than 100m

- Or –

“*I was happy*”
“*I enjoyed life*”
Local dependency

- Yen’s Q3
- Correlation between the item residuals
- Cut off +.2 is indicative of local dependency

NB – this is a simplified representation, correlations assessed between all items
Local dependency

- Yen’s Q3
- Correlation between the item residuals
- Cut off +.2 is indicative of local dependency

Item two = “Are you able to work?”, Item three = “How would you rate your ability to work?”
Local dependency

- Yen’s Q3
- Correlation between the item residuals
- Cut off +.2 is indicative of local dependency

\[
\begin{align*}
r &= .01 \\
r &= .30 \\
r &= .01 \\
r &= -.30
\end{align*}
\]

- Item one
- Item two
- Item three
- Item four
- Item five
Local dependency

- Yen’s Q3
- Correlation between the item residuals
- Cut of off +.2 is indicative of local dependency

<table>
<thead>
<tr>
<th>Item</th>
<th>R</th>
<th>Item</th>
<th>R</th>
<th>Item</th>
<th>R</th>
<th>Item</th>
<th>R</th>
<th>Item</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>one</td>
<td>.01</td>
<td>two</td>
<td>.30</td>
<td>three</td>
<td>.01</td>
<td>four</td>
<td>-.30</td>
<td>five</td>
<td></td>
</tr>
</tbody>
</table>
Differential item functioning

- Some items introduce bias by measuring demographic differences between test-takers
Differential item functioning

- Functional ability measure
- “Help with eating, dressing and using the toilet”
- Higher scores in Turkey

Scott et al., 2006. Comparing translations of the EORTC QLQ-C30 using differential item functioning analysis
**Unidimensionality**

- Measures are always uninterpretable if they combine more than one dimension.

- A measure that included height and weight together would be useless for assessing either height or weight.

- Some IRT techniques (multidimensional item response theory) can deal with multidimensionality, but are theoretically complex.

- For most IRT models, we must be certain of unidimensionality.

- Mokken analysis is a good alternative to factor analysis for establishing unidimensionality and scalability.
Reliability

- Coefficient Alpha / Marginal reliability
  - Average for the whole scale
  - Not matched to your population
  - Reliability around a cut-off might be low, for example

With IRT, we can do better!

Item information, standard error and reliability are all related
Individual item information

- **Blue item** “Sometimes I feel a bit sad”
- **Red item** “I often feel suicidal”
Individual item information

Item Information Curves

Information

Theta $\theta$

-4 -2 0 2 4
Individual item information

Item Information Curves

Clear, easily interpreted item

Unclear item
Individual item information

Clear, easily interpreted item
“I feel depressed”

Unclear item
“I get butterflies in my tummy”
Test information

\[ SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \]
Information and **Standard Error**

\[ SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \]

- Error of measurement inversely related to information
- Standard error (SE) is an estimate of measurement precision at a given theta
- Reliability can be calculated from SE and information
Information and Reliability

\[
\text{Reliability} (\theta) = 1 - \left( \frac{1}{\sqrt{I(\theta)}} \right)^2
\]
IRT parameters (1pl)

\[
P(1|\theta) = \frac{e^{\theta_n - bi}}{1 + e^{\theta_n - bi}}
\]

\begin{tabular}{|c|c|}
\hline
\(a\) & \(b\) \\
\hline
1 & -1 \\
1 & -0.05 \\
1 & 0 \\
1 & 0.05 \\
1 & 1 \\
\hline
\end{tabular}
IRT parameters (2pl)

\[ P(1|\theta) = \frac{e^{a_i(\theta_n - b_i)}}{1 + e^{a_i\theta_n - b_i}} \]
IRT parameters (1pl polytomous)

<table>
<thead>
<tr>
<th>a</th>
<th>b1</th>
<th>b2</th>
<th>b3</th>
<th>b4</th>
<th>b5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-3.75</td>
<td>-3</td>
<td>-1</td>
<td>0.25</td>
<td>1.5</td>
</tr>
</tbody>
</table>
What is computer adaptive testing?

- Computerised method for administering items that ‘learns’ from participant responses and usually* administers items based on the degree of information it gives us about the test-taker

  * = Can be overridden.

- Technique for maximising information about each candidate, whilst minimising the length of the assessment
Why use CAT?

- Compared to paper-based tests it is-
  - More flexible
  - More efficient
  - More accurate
  - Better targeted

- Integrated feedback
- Less ‘gaming’
Why use CAT?

• Shorter assessments
• Avoid asking bright candidates easy items
• Avoid asking distressing items to people with low levels of a construct (e.g., functional impairment)
• Calibrate item banks that cover a wide range of the construct (e.g., cognitive impairment)
• Avoid gaming on repeated measures

• Electronic assessment
• Integrated feedback
Administer item

Record score

Estimate theta

Is stopping rule met

Select next item

Calculate score

Give feedback

N

Y
Standard questionnaire to assess Lisa

\[
\begin{align*}
\star &= \text{A question from our questionnaire} \\
\text{Maths ability} &= \frac{2+2}{1134 \times 16}
\end{align*}
\]
Standard questionnaire to assess Lisa

= A question from our questionnaire

Maths ability

Lisa’s ability
A (really) simple introduction to CAT

= A question from our questionnaire

Maths ability
Computer adaptive test to assess Lisa

= A question from our questionnaire

Maths ability

8 x 4
Computer adaptive test to assess Lisa

= A question from our questionnaire

Maths ability
Computer adaptive test to assess Lisa

= A question from our questionnaire

Maths ability

182 + 427
Computer adaptive test to assess Lisa

= A question from our questionnaire

Maths ability

★ ★
Computer adaptive test to assess Lisa

= A question from our questionnaire

Maths ability

1134 x 16
Computer adaptive test to assess Lisa

= A question from our questionnaire

Maths ability
Computer adaptive test to assess Lisa

= A question from our questionnaire

Maths ability

1712 + 3218
Computer adaptive test to assess Lisa

= A question from our questionnaire

Maths ability

⭐️⭐️⭐️⭐️⭐️
Computer adaptive test to assess Lisa

= A question from our questionnaire

Maths ability

204 x 16
Computer adaptive test to assess Lisa

= A question from our questionnaire

Maths ability
Computer adaptive test to assess Lisa

= A question from our questionnaire

Maths ability

Lisa’s ability
What does a CAT know?

Score input

Item characteristic curves

Item information

[ Next item, theta estimate, standard error ]
Item Information

- Rasch model
Item Information

• Graded Response Model
Item Selection

• Maximum information at estimated level of theta

• Bayes Modal (MAP)
• Expected *a posteriori* (EAP)
Maximium Information

- Maximum information at estimated level of theta
- Proposed by Fisher, developed by Lord
- Each item is selected to provide the maximum information, given the provision estimate of ability (theta)
- Generally items with the steepest discrimination
- Most efficient way to run a test
Bayes Modal (MAP)

- Maximum *a posteriori*
- Bayesian technique for item selection
- Bayesian, so it takes into account the distribution of the population
- Essentially it is MI * population distribution
Bayes Modal (MAP)

- Maximum *a posteriori*
- Bayesian technique for item selection
- Bayesian, so it takes into account the distribution of the population
- Essentially it is MI * population distribution

MAP

![Bayesian adjusted score graph](image_url)
Expected *a posteriori* (EAP)

- Instead of taking maximum point of the Bayesian adjusted likelihood function, we take an average value weighted by the EAP function.

![Bayesian adjusted score graph](image)
Specific Information

- Proposed by Davey and Fan
- Administer items to achieve pre-selected information targets
- Useful for tests with a cut-off where we don’t ‘care’ about the information relating to people who are far away from the cut-off
Stopping rules

- Test length (*e.g.*, 20 items, 15 items)

- Test time (5 minutes)

- Reliability of theta estimate (*standard error*)

- Other, clever stuff
Reliability and **Standard Error**

- $\text{Alpha}(0.90) = \text{SE}(0.32)$
- $\text{Alpha}(0.80) = \text{SE}(0.45)$
- $\text{Alpha}(0.70) = \text{SE}(0.55)$
Item Selection quiz..

- Theta = 0
- Theta = -1.8
- Theta = 3
- Theta = -2.2

- What additional information would a Bayesian want?

- Which item would you never use?
Item Selection quiz..

- $\Theta = 0$
- $\Theta = -1.8$
- $\Theta = 3$
- $\Theta = -2.2$
Exercise

- Item information
- Item difficulty
- Make a CAT in your group.