Introduction to Item Response Theory

Prof John Rust, j.rust@jbs.cam.ac.uk
David Stillwell, ds617@cam.ac.uk
Aiden Loe, bsl28@cam.ac.uk
Luning Sun, ls523@cam.ac.uk

www.psychometrics.cam.ac.uk
Goals

• Build your own basic tests using Concerto

• General understanding of CTT, IRT and CAT concepts
  • No equations!
Understanding the Latent Variable
What is a construct?

- 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
Latent variable as the cause of item values

- 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
Classical Measurement Assumptions

\[ X = T + e \]

- \( X = \text{observed score} \)
- \( T = \text{true score} \)
- \( e = \text{error} \)
Classical Measurement Assumptions

1. Items’ means unaffected by error if have a large number of respondents

2. One item’s error is *not* correlated with another item’s error

3. Error terms are *not* correlated with the true score of the latent variable
Classical Test Theory

- **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 vs. Latent Trait Models

- Test level is the basis for CTT (not the item).
- While the statistics outcome are often extended to similar students taking a similar test; they only really apply to those students taking that test (norm based).
- Latent trait models go further and investigate the underlying traits which are producing the test performance.
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)
Items Analysis

Classical Test Theory

Latent Trait Models

Item Response Theory

Rasch Models

1P  2P  3P  4P

Similar
Item Response Theory

- Item Response Theory (IRT) – refers to a family of latent trait models.
- They are used to establish psychometric properties of items and scales
- IRT has many advantages over CTT that have brought IRT into more frequent use
3 Basics Components of IRT

- Item Response Function (IRF) – Mathematical function that relates the latent trait to the probability of endorsing an item
- Item Information Function – an indication of item quality; an item’s ability to differentiate among respondents
- Invariance – position on the latent trait can be estimated by any items with known IRFs. The ICCs are sample independent (If you spilt by groups, the item parameters should be the same).
Probability of getting item right

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)
Rasch vs IRT philosophy
(roughly)

- Rasch – Model primacy. Data should fit the model

- IRT – Data primacy. Model should fit the data

Different fields, constructs, and purposes demand different approaches.

- The jury is out as to whether this has any impact on measurement in the real world.
5 items of varying difficulty (b) and discrimination (a)

Two-Parameter Logistic Model (2PL)
One item showing the guessing parameter (c)

Three-Parameter Model (3PL)
Option Response Function

Probability of Correct + Probability of Incorrect = 1

Binary items

Correct response

Incorrect response

Theta
Graded Model
(example of a model with polytomous items – e.g. Likert Scales)

Category Response Curves for an item representing the probability of responding in a particular category conditional on trait level

“I experience dizziness when I first wake up in the morning”
(0) “never”
(1) “rarely”
(2) “some of the time”
(3) “most of the time”
(4) “almost always”
Fisher Information Function
Test Information Function

- Test Information Function (TIF) – The IIFs are additive so that we can judge the test as a whole and see at which part of the trait range it is working the best.
(Fisher) Test Information Function

Probability

Theta

Three items

*Not to scale
TIF and Standard Error (SE)

- Error of measurement inversely related to information
- Standard error (SE) is an estimate of measurement precision at a given theta
Scoring

Test:
1. Normal distribution
2. q1 – Correct
3. q2 – Correct
4. q3 - Incorrect
Quiz!
Quiz!

- Which of the following is not an IRT parameter?
  - Unfolding
  - Discrimination
  - Difficulty
  - Probability
  - Guessing

- [HARD ONE!] Name the parameters from their shorthand
  a - 
  b - 
  c -
Quiz!

• Which of the following is not an IRT parameter?
  - Unfolding
  - Discrimination
  - Difficulty
  - Probability
  - Guessing

• [HARD ONE!] Name the parameters from their shorthand
  a - discrimination
  b – difficulty
  c - guessing
IRT assumptions

• Unidimensionality – All items are assumed to measure a single common factor.

• Local independence – item responses are uncorrelated after controlling for the latent trait.

• Homogenous population – No differential item functioning (DIF) between groups.
# 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>
IRT in R

ltm package

Suggested Resource:

Dr Philipp Doebler of the University of Munster describes the latest thinking on adaptivity in psychometric testing to an audience of psychologists.

The dimension of interest is women’s mobility and social freedom.


Data is available within R software package “ltm”
“Mobility” Survey

Women were asked whether they could engage in the following activities alone (1 = yes, 0 = no):

1. Go to any part of the village/town/city.
2. Go outside the village/town/city.
3. Talk to a man you do not know.
4. Go to a cinema/cultural show.
5. Go shopping.
6. Go to a cooperative/mothers' club/other club.
7. Attend a political meeting.
8. Go to a health centre/hospital.
Ltm package

install.packages("ltm")
require(ltm)
help(ltm)
head(Mobility)
my1pl<-rasch(Mobility)
my1pl
summary(my1pl)
plot(my1pl, type = "ICC")
plot(my1pl, type = "IIC", items=0)
## Ltm package

### rasch

myrasch <- rasch(Mobility, cbind(9,1))

my2pl <- ltm(Mobility ~ z1)

anova(my1pl, my2pl)

(the smaller the better!)

Now plot ICC and IIC for 2pl model.
ltm package – scoring

```r
resp<-matrix(c(1,1,1,0,1,0,1), nrow=1)
factor.scores(my2pl, method="EAP", resp.patterns=resp)

EXPLAIN: "$" addressing

theta = dataCAT$score.dat$z1
sem = dataCAT$score.dat$se.z1

mobIRT <- factor.scores(my2pl, resp.patterns=Mobility, method="EAP")
head(mobIRT$score.dat)
```
Compare IRT and CTT scores

CTT_scores <- rowSums(Mobility)
IRT_scores <- mobIRT$score.dat$z1
plot(IRT_scores, CTT_scores)

#Plot the standard error and scores
IRT_errors <- mobIRT$score.dat$se.z1
plot(IRT_scores, IRT_errors, type="p")
Why use Item Response Theory?

- Reliability for each examinee / latent trait level
- Modelling on the item 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)
- Score is item independent
- 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
Computerized Adaptive Testing

- 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

Example of CAT

![Graph showing normal distribution with incorrect and correct response probabilities and theta values.](image)

- Incorrect response
- Correct response
- Normal distribution
- Difficulty
- Most likely score
Elements of CAT

- 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
CAT Demos

http://planning.e-psychoometrics.com/test/cat-demo

http://planning.e-psychoometrics.com/test/cat-dna

Demos created by Aiden Loe

Suggested Resource:
Dr Philipp Doebler of the University of Munster describes the latest thinking on adaptivity in psychometric testing to an audience of psychologists.