Item Response Theory and Computerized Adaptive Testing

Hands-on Workshop, day 2
John Rust, jnr24@cam.ac.uk
Iva Cek, ic274@cam.ac.uk
Luning Sun, ls523@cam.ac.uk
Michal Kosinski, mk583@cam.ac.uk
Goals

* General understanding of IRT and CAT concepts
  * No equations!
* Acquire necessary technical skills (R)
* Tomorrow: Build your own IRT-based CAT tests using Concerto
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)
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
Probability of getting item right

Ratio of correct responses to items on different level of total score
Please mind that those and many other graphs presented here are just Excel based mock-ups created for the presentation purposes rather than representing actual data.

**Item Response Function**

**Binary items**

- Probability of getting item right
- Measured concept (theta)

**Parameters:**
- Difficulty
- Discrimination
- Guessing
- Inattention

**Models:**
- 1 Parameter
- 2 Parameter
- 3 Parameter
- 4 Parameter
- unfolding
One-Parameter Logistic Model/Rasch Model (1PL)

7 items of varying difficulty (b)
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: $b=0.0$, $a=1.0$, $c=0.2$
Option Response Function

Binary items

Probability of Correct + Probability of Incorrect = 1
Graded Model
(example of a model with polytomous items – e.g. Likert Scales)

“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”

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

![Graph showing Fisher Information Function with Probability on the y-axis and Theta on the x-axis.](graph.png)
(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>Modelling / Interpretation</th>
<th>Classical</th>
<th>IRT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total score</td>
<td>Individual items (questions)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Adaptivity</th>
<th>Virtually not possible</th>
<th>Possible</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Score</th>
<th>Depends on the items</th>
<th>Item independent</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Item Parameters</th>
<th>Sample dependent</th>
<th>Sample independent</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Preferred items</th>
<th>Average difficulty</th>
<th>Any difficulty</th>
</tr>
</thead>
</table>
Why use Item Response Theory?

- Reliability for each examinee / latent trait level
- Modelling on the item level
- Examinee / Item parameters on the same scale
- Examinee / Item parameters invariance
- Score is item independent
- Adaptive testing
- Also, test development is: cheaper and faster!
IRT in R

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”
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.
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

resp<-matrix(c(1,1,1,1,0,1,0,1), 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")
Model FIT

Checking model fit:
- `margins(my1pl)`
- `GoF.rasch(my1pl, B=199)`
Introduction to CAT
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
Example of CAT

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
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
Classic approaches to item selection

- Maximum Fisher information (MFI)
  - Obtain a current ability estimate
  - Select next item that maximizes information around the current ability estimate
- Urry’s method (in 1PL equals MFI)
  - Obtain a current ability estimate
  - Select next item with a difficulty closest to the current one
- Other methods:
  - Minimum expected posterior variance (MEPV)
  - Maximum likelihood weighted information (MLWI)
  - Maximum posterior weighted information (MPWI)
  - Maximum expected information (MEI)
Examples of item overexposure prevention

* Randomesque approach (Kingsbury & Zara, 1989)
  * Select >1 next best item
  * Randomly choose from this set
* Embargo on overexposed items
* Location / Name / IP address rules

Content Balancing

* Ascertain that all subgroups of items are used equally
catR package

CAT in R

Suggested Resource:

Dr Philipp Doebler of the University of Munster describes the latest thinking on adaptivity in psychometric testing to an audience of psychologists.
install.packages("catR")
require(catR)
c<-coef(my2pl)
itemBank <- cbind(c[,2], c[,1], 0, 1)
catBank<-createItemBank(itemBank, model="2pl")
catBank
catBank$itemPar
plot(catBank$infoTab[,1])
plot(my2pl, type = "IIC", items=1)
Choose the item to start with:

* max info around average?
  
  ```r
  plot(my2pl, type = "IIC")
  plot(my2pl, type = "IIC", items=4)
  ```

* Random one?
catR

items_administered <- c(4)
responses <- c(1)

it <- itemBank[items_administered, 1:4, drop=F ]
theta <- thetaEst(it, responses)

q <- nextItem(catBank, theta, out=items_administered)
q$item
\[ p(\theta) = guessing + (1 - guessing) \frac{1}{1 + e^{-\text{discrimination}(\theta - \text{difficulty})}} \]

Assumption of Local Independence – A response to a question is independent of responses to other questions in a scale after controlling for the latent trait (construct) measured by the scale.

Assumption of Unidimensionality - the set of questions are measuring a single continuous latent variable (construct).

under the assumption of a normal \( \theta \) distribution, to the biserial item-test correlation \( \rho \) (Linden & Hambleton, 1997). For item i the relationship is:

Difficulty = \( p / \sqrt{1-p} \)

\[ I = \text{discrimination}^2 p(\text{correct}) * (p(\text{incorrect}) \ (2pl)) \]

Standard error = \( 1 / \sqrt{\text{information}} \)