UNIVERSITY OF CAMBRIDGE

The Psychometrics Centre

Summer School in Applied Psychometric Principles

Peterhouse College

13th to 17th September 2010

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

This course is prepared by

Anna Brown (University of Cambridge)

Jan Böhnke (University of Trier)

Tim Croudace (University of Cambridge)

Introductions

- Your name
- Your background
- Your field of research
- Your needs and expectations from this course

Programme

- Day 1: Introducing Item Response Theory models (binary).
- Day 2: Two- and three-parameter IRT models. Introducing models for polytomous data. Test information in IRT and reliability. Testing assumptions and assessing model fit.
- Day 3: The Rasch model for both binary and polytomous data. Properties of Rasch measurement and scaling.
- Day 4: Introducing concepts of measurement invariance. Investigating Differential Item Functioning (DIF) using various approaches (Mantel-Haenszel and Confirmatory Factor Analysis (CFA) with covariates).
- Day 5: Example applications of Item Response Theory: test equating and Computer Adaptive Testing (CAT).

Daily schedule

- Monday 1.00 pm Lunch
  2.00 pm - 5.00 pm
- Tuesday - Thursday
  9.00 am - 5.00 pm
  1.00 pm Lunch
- Friday 9.00 am - 1.00 pm
  1.00 pm Lunch

Introducing Item Response Theory models (binary)

Day 1

Anna Brown, PhD
University of Cambridge

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

References

- Hambleton, Swaminathan & Rogers (1991). Fundamentals of Item Response Theory.
- Embretson, S. & Reise, S. (2000). Item Response Theory for psychologists.
- R.J. de Ayala (2009). The theory and practice of Item Response Theory.
- Van der Linden, W. & Hambleton, R. eds. (1997). Handbook of modern Item Response Theory.
- McDonald, R. (1999). Test theory.

Tests are not perfect measurements

- Psychometric tests are certainly different from measurements we routinely use every day – such as temperature, weight, length etc.
- Test should be viewed as a series of small experiments outcomes of which are recorded
  - from which a measure is inferred (van der Linden & Hambleton).
  - Ways to cope with experimental error is 1) matching or standardisation, 2) randomisation, 3) statistical adjustment.

Classical Test Theory

- The classical test model

$$X = T + E$$

  - X = test score (observed)
  - T = true score – defined as expected test score (unobserved)
  - E = random error (unobserved)
- No constraints are imposed on X thus the model always holds
- No distributional assumptions about X, T, or even E need to be made (in which case equation has no solution)

CTT Assumptions:

1. $\overline{E} = 0 \qquad E(X) = T$
2. $\rho_{TE} = 0$
3. $\rho(E_1, E_2) = 0$

Definition of Parallel Tests:

Two or more tests measuring the same content and

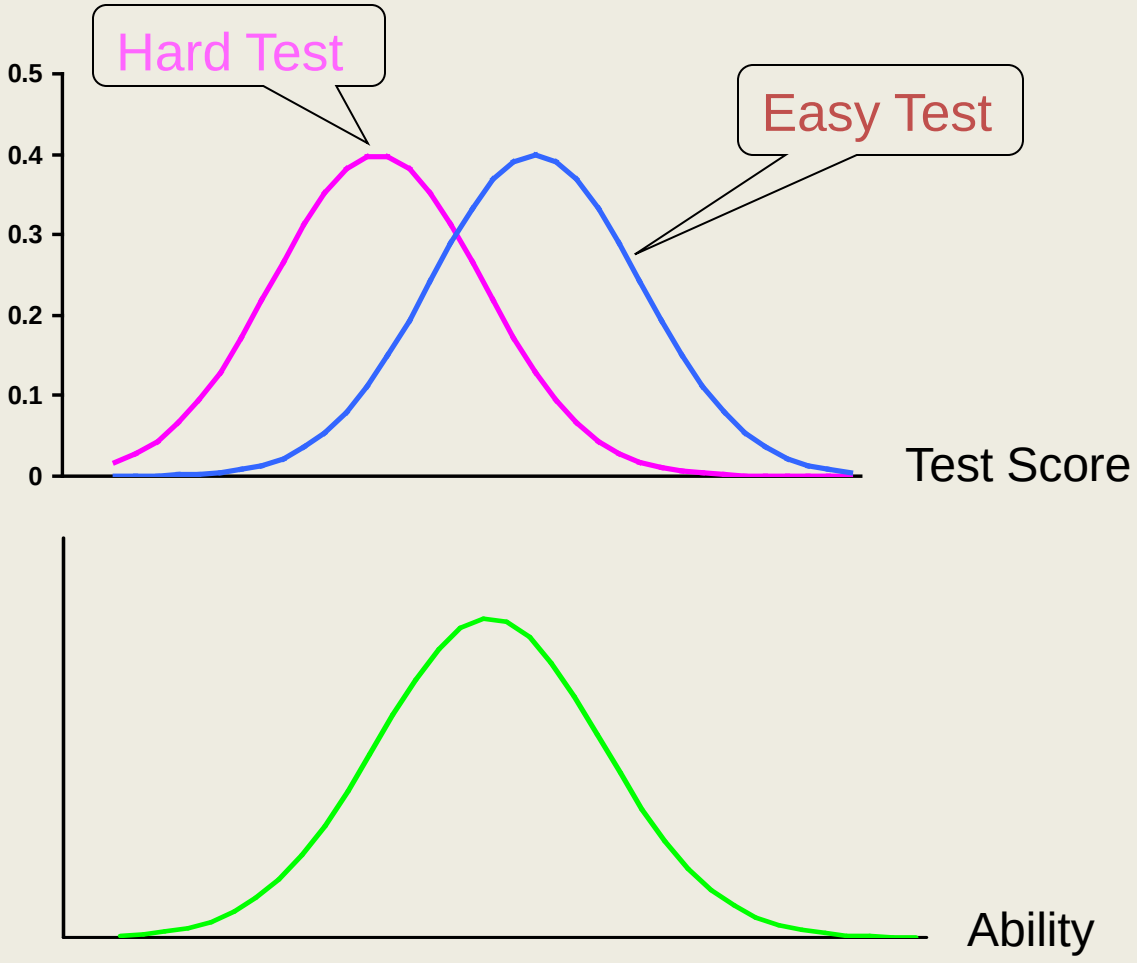
1. $T_1 = T_2$
2. $\sigma^2(E_1) = \sigma^2(E_2)$

- CTT model is based on weak assumptions (that are easy to achieve assumptions with many test data sets); therefore, CTT has wide applicability in the testing field!

True scores are test dependent

- In CTT, true score is fully determined by the test as designed
  - not by some "state" inside the examinee that is independent of test
- True score only has meaning conditional on standardised error variables
- Specifics of a particular testing situation, e.g. properties of test items are nuisance error variables that escape standardisation
- Statistical adjustment is needed to control for these nuisance factors

Test score and ability distribution



Limitations of Classical Test Theory

- Examinee proficiency scores are item dependent.
- Item statistics are sample dependent.
- The common estimate of measurement error (SEm) is group-based.
- Modeling of data is at the test score level (X=T+E) but item level modeling is needed for flexibility of use
  - item banks
  - computer-adaptive tests
  - improved score reporting, and more…

What do test developers want?

- Examinee parameter invariance
- Item parameter invariance
- Estimate of error for each examinee
- Modeling examinee responses at the item level for flexibility in test item selection
- Examinees and items on a common reporting scale (optimal test design)

Item Response Theory (IRT)

- Models to make statistical adjustments in test scores have been developed in IRT
  - Adjustments for such item properties as difficulty, discriminating power, and liability to guessing.
- IRT models the test behaviour not at the arbitrary test score level, but at the item level

History of IRT

- Can be traced to the 1940s (work by Lawley, Richardson, Tucker).
- 1950s - Lord, Birnbaum, and Rasch.
- 1960s and 1970s - work by Bock, Lord, McDonald, Samejima, Rasch, Fischer, Wright, Andrich, Goldstein, and many more.
- Interest in computer adaptive testing was a major force in the development in the 1960s (but there was no computer power).
- With software, the IRT field has developed rapidly.

Item Response Theory

- IRT (also *latent trait theory*) is a model-based measurement in which trait level estimates depend on both person's responses and item properties.
  - Links between traits (*what the test measures, and what is of interest to the test designer*) and item responses are made through non-linear models that are based upon assumptions that can always be checked.

The latent trait

Notation: "**theta**" $\theta \in (-\infty, +\infty)$

- The latent trait is simply the label used to describe what the set of test items (tasks) measures. [*Has been common to say "ability" or "proficiency" regardless of what the test measures*.]
- Latent trait can be **broadly or narrowly defined** psychomotor, aptitude, achievement or psychological variable.
- No reason to think of trait or "ability" as fixed over time. In fact, it **should be** influenced by instruction, training, aging…
- Validation studies are required to determine what a test measures—content, criterion-related, and construct evidence.

The item responses

Notation: u_{ij} – response of examinee j to item i

- Test items most often assume categorical response
- Ability tests typically produce *binary* responses (correct – incorrect), for example, $u_{ij}=1$ if correct and $u_{ij}=0$ incorrect
  - Sometimes choice alternatives can be modelled directly using *nominal* categories
- Questionnaires that employ *rating scales* most often have ordered categorical (*ordinal*) responses
  - Might have 3, 4, 5, 7 or even 9 rating categories
  - Rating scales can be symmetrical (agree-disagree) and not (never-always)

The item parameters

Notation: "**a**", "**b**", "**c**" and others

e.g. discrimination $a_i \in (0, +\infty)$

and difficulty $b_i \in (-\infty, +\infty)$

- Simply symbols at this point – meaning will depend on the model
- Vary in different IRT models depending on which item properties are assumed to influence the probability of item responses

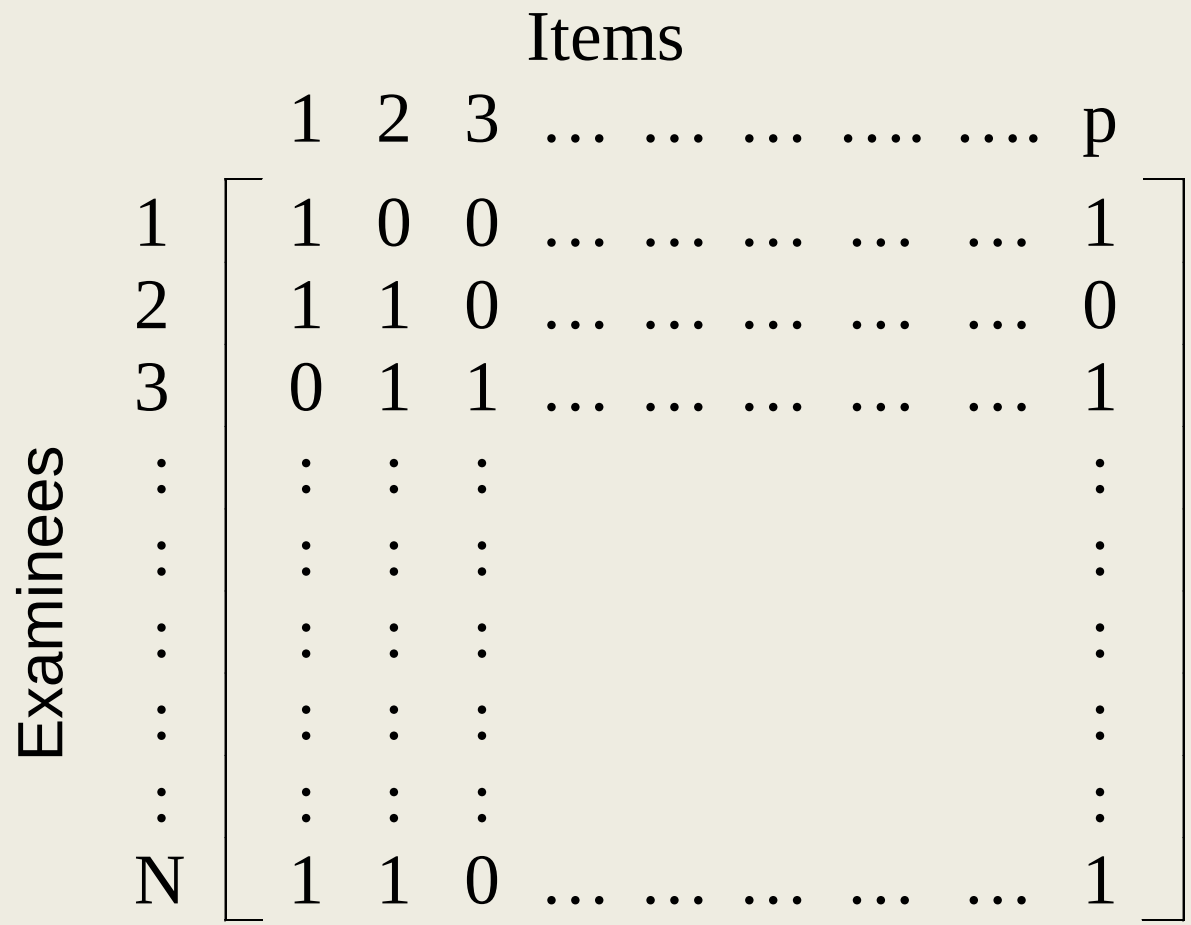
Introduction to IRT

ITEM RESPONSES AS FUNCTIONS OF THE LATENT TRAIT

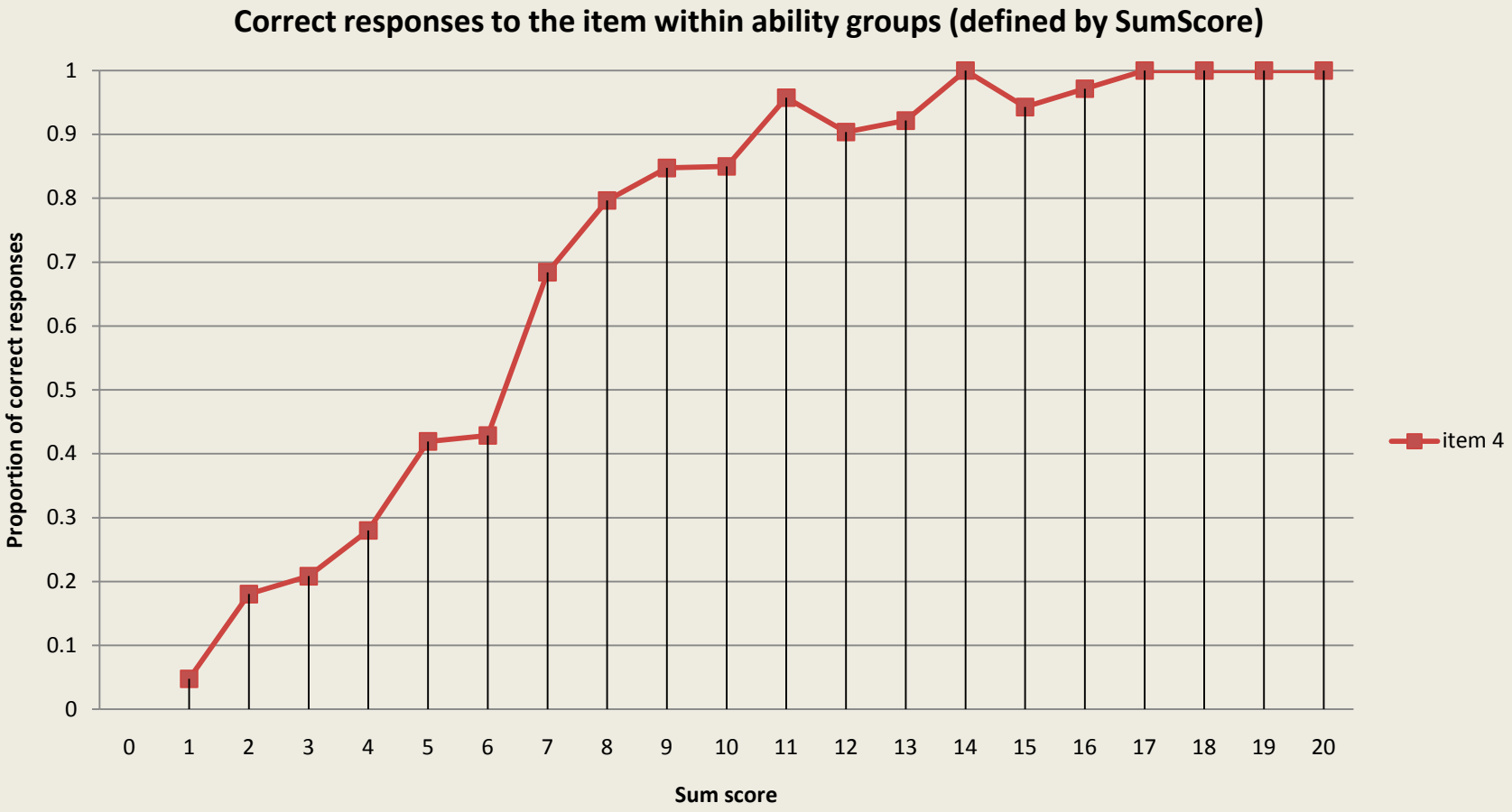
Example ability test

- Consider a test with 20 items.
- Each item is assumed to 'sample' one underlying (latent) dimensions of 'achievement' or 'ability', say aptitude for mathematics.
- Administered to 1000 examinees.
- Let's start with counting items that were answered correctly for each examinee (*sum score* or number correct).
- Use the sum score as a proxy for mathematical ability.

Binary test data

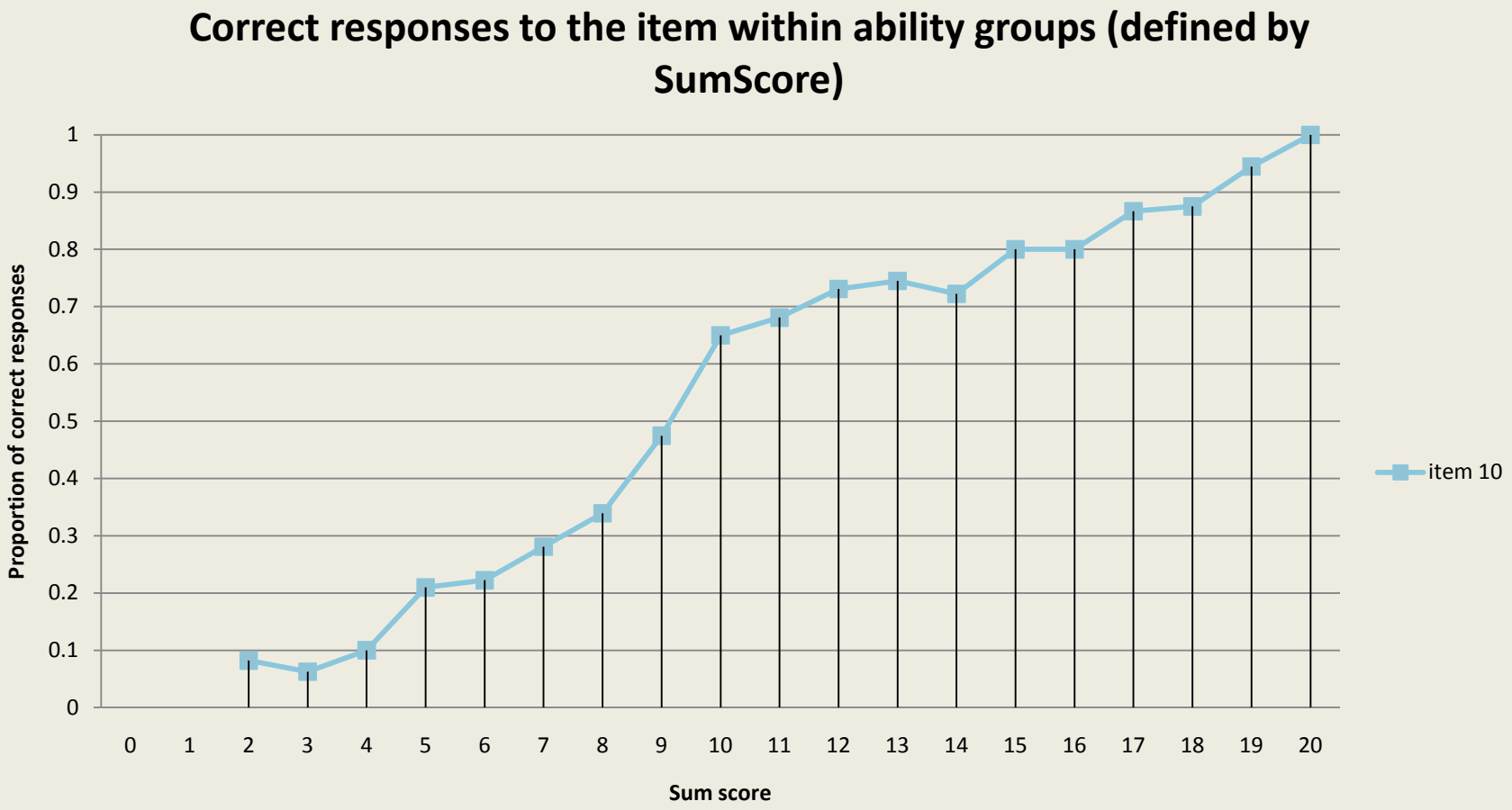


Likelihood of correct response as function of ability

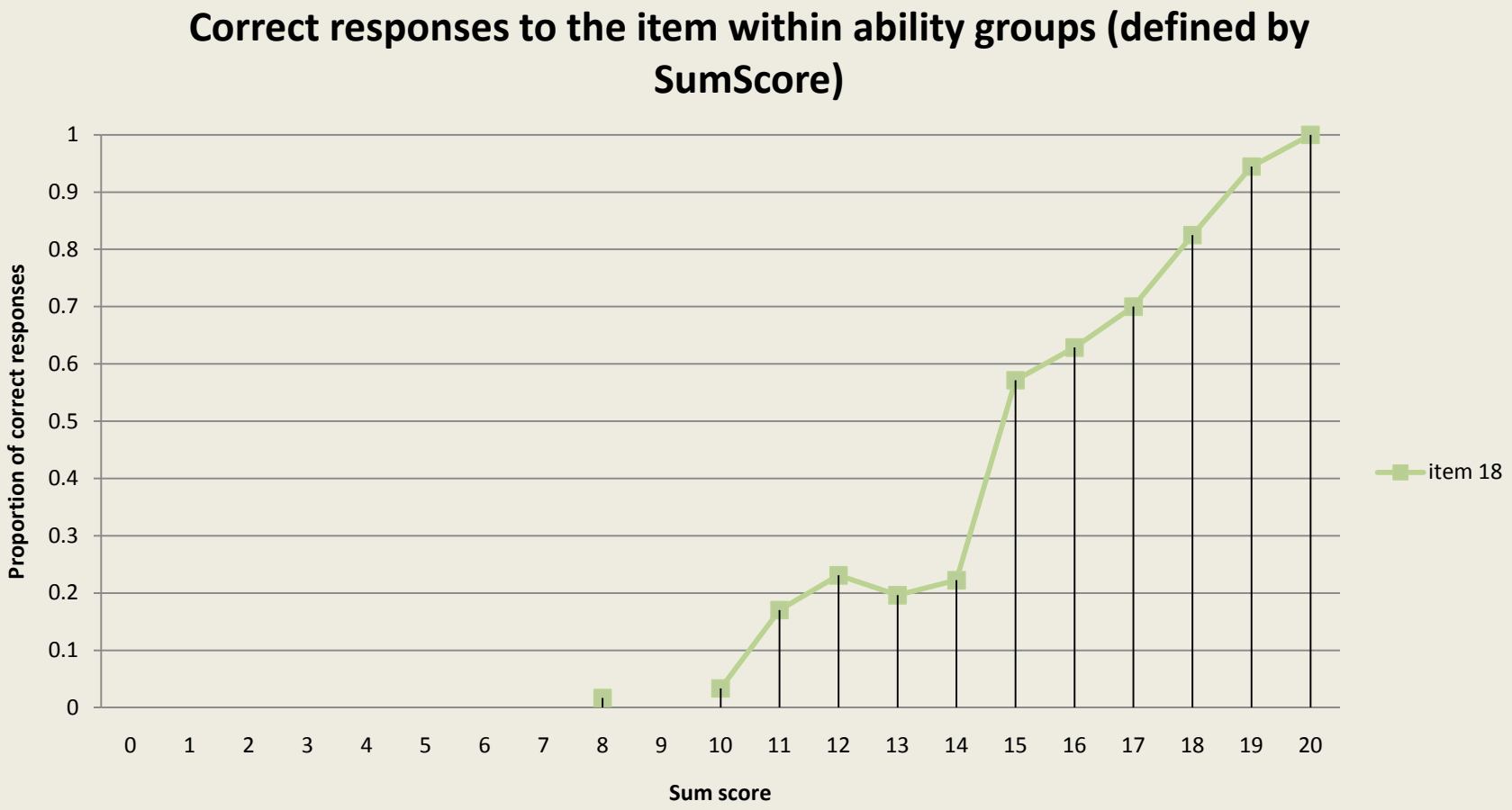


UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

…and for another item

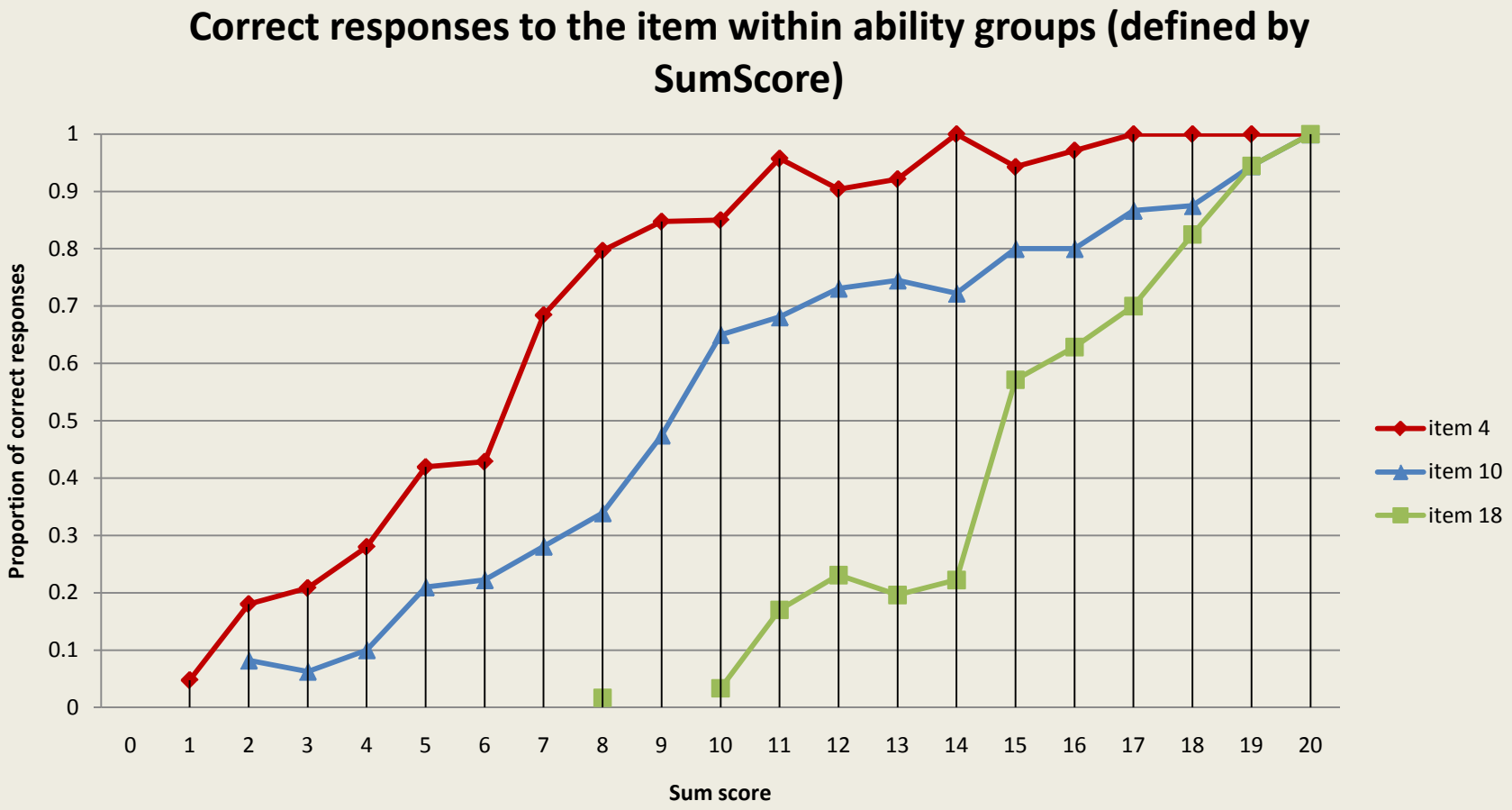


…and one more item



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

What can be said about these items?



Item Response Function (IRF)

Notation: $P_i(u_{ij} = 1 \mid \theta)$ $P_i(\theta) \in (0, 1)$

- Called Item Response Function (IRF)
  - or Item Characteristic Curve (ICC) – less appropriate in multidimensional case
- Links the probability of an item response to the latent trait
- In this ability example (and in many other IRT applications), probability of a correct response should increase monotonically as ability increases
- Has to be bounded between 0 and 1
  - Cannot be a linear function of ability!

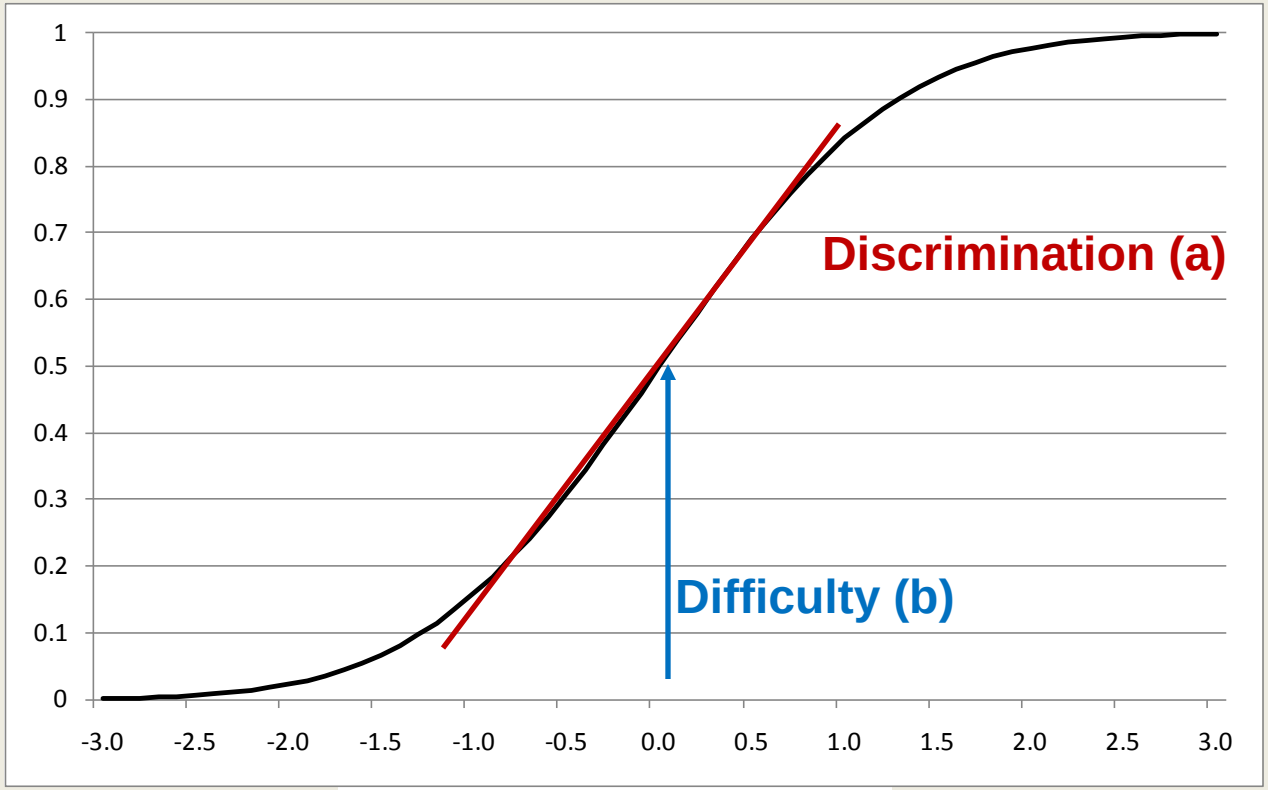
Normal-ogive model

$$P_i(\theta) = \Phi\left(a_i(\theta - b_i)\right) = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

- Familiar *cumulative normal distribution* function with 2 item parameters (can be looked up in tables)
- The first ever IRT model. The first coherent treatment was given by Lord (1952)
- Lord and Novick (1968) showed that under normal ability distribution, parameters a and b are related to CTT difficulty and item-test correlation
- Maths is horrible so models with logistic links eventually became more popular (though their IRFs are virtually indistinguishable)

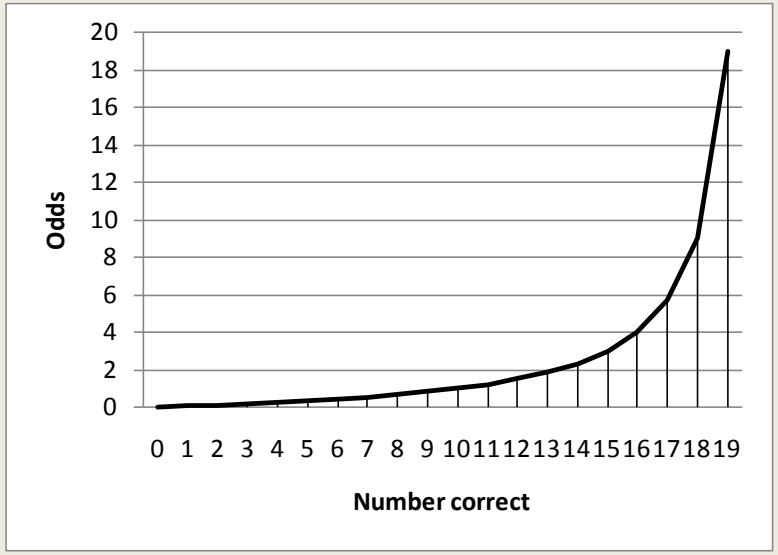
Example of normal-ogive IRF

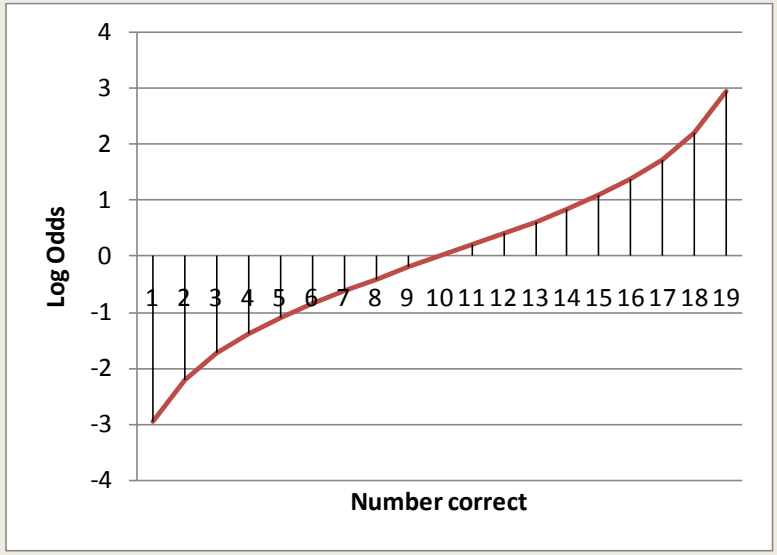
- With parameters a=1, b=0



Odds and log odds

- Odds = ratio of the number of successes to the number of failures $P/(1-P)$
  - In a test with 20 binary items the odds are distributed as follows:





The Rasch model

- In 1950th Rasch proposed a simple relationship between the person's trait score and item difficulty for describing odds of passing an item

$$\ln[P/(1-P)] = \theta - b$$

- Same interpretation of the difficulty parameter as in the normal-ogive model – point on the scale where probabilities of success and failure are equal

$$P(u_i = 1 \mid \theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

- Logistic link function, and maths is easy (though IRF is virtually indistinguishable from normal-ogive)

Birnbaum's logistic models

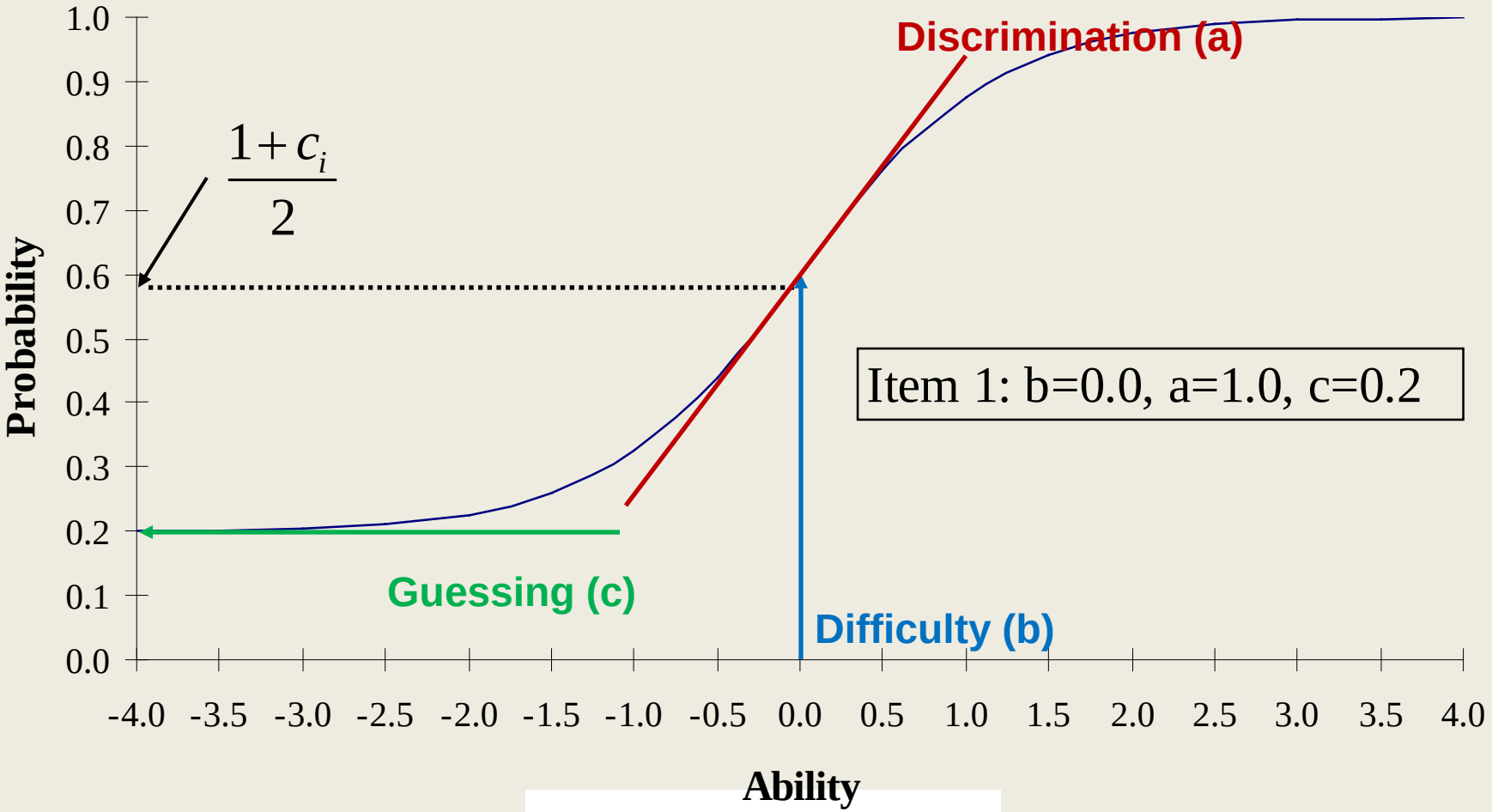
- Worked in late 1950th- main motivation was to make the work begun by Lord statistically feasible
- Proposed to replace the normal-ogive by the logistic model
  - Based on Haley (1952) result: |N(x)-L(1.7x) |<0.01

$$P\left(u_i = 1 \mid \theta\right) = \frac{e^{Da_i\left(\theta - b_i\right)}}{1 + e^{Da_i\left(\theta - b_i\right)}}$$

- Also proposed a third parameter to account for guessing

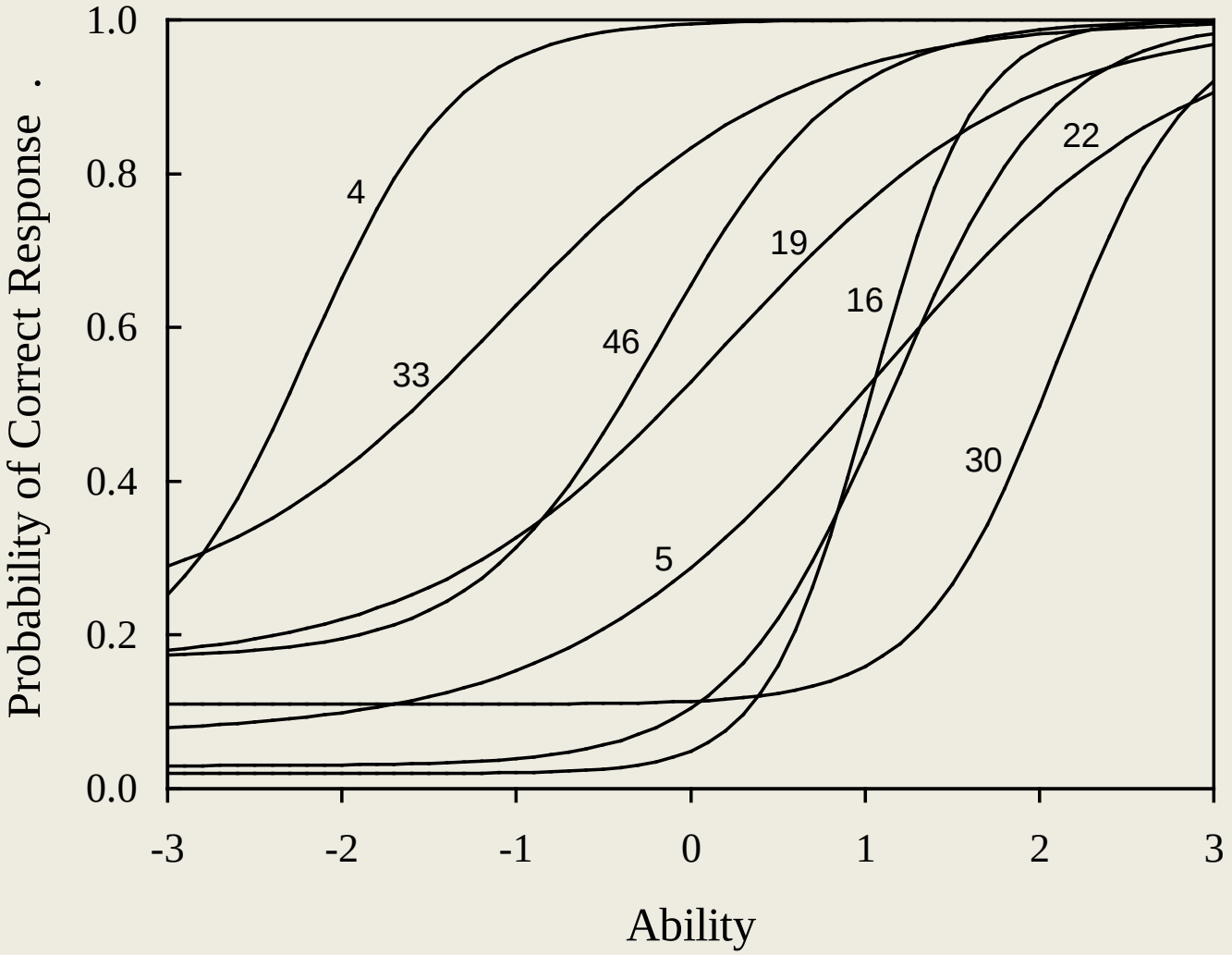
$$P\left(u_i = 1 \mid \theta\right) = c_i + \left(1 - c_i\right)\frac{e^{Da_i\left(\theta - b_i\right)}}{1 + e^{Da_i\left(\theta - b_i\right)}}$$

Item Parameter interpretations



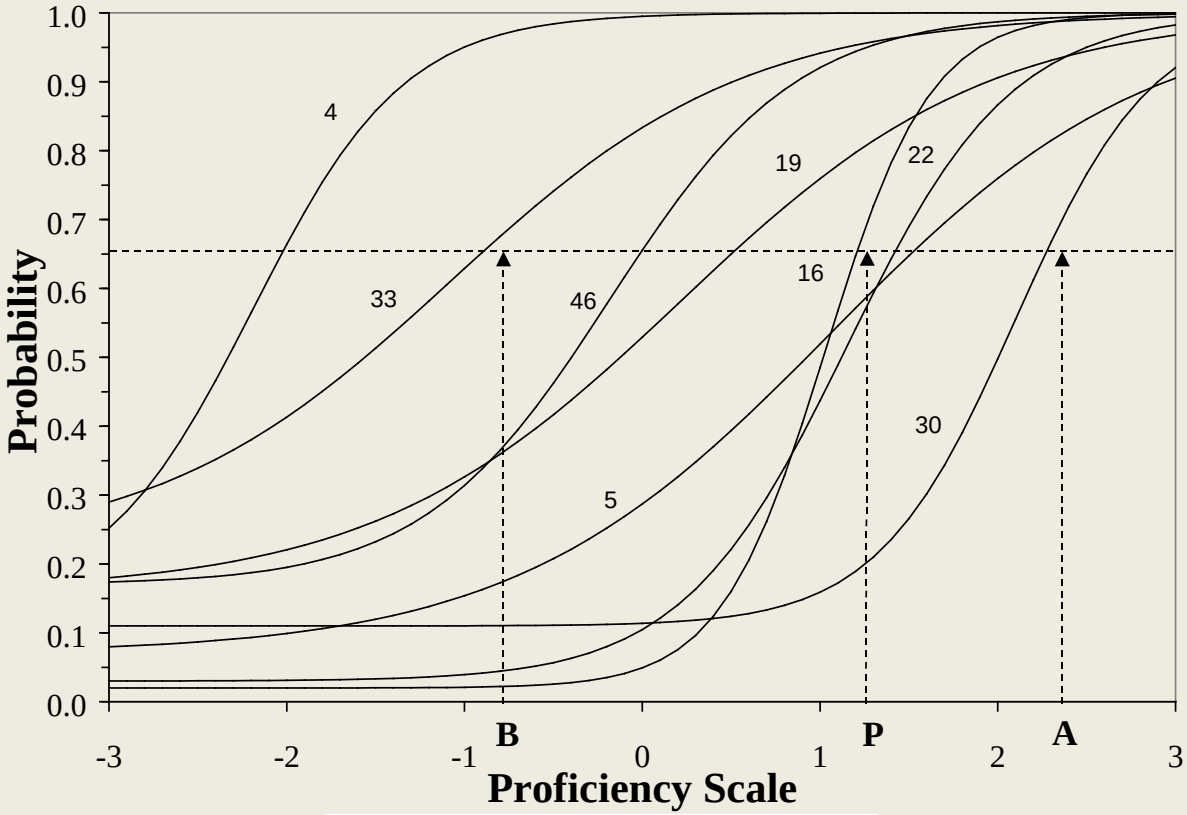
UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Examples of eight IRFs



Item mapping and benchmarking

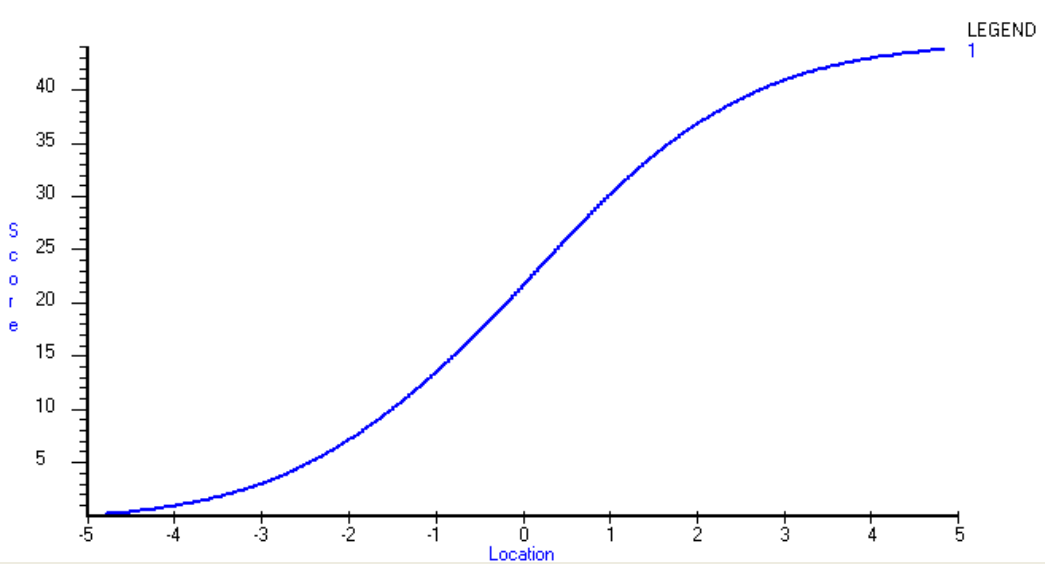
- In IRT items and examinees are on the same scale



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Test response function

- Adding all item response functions (probability of response =1) will produce the test information function
- It predicts relationships between sum score and the IRT estimated score
  - This relationship is not linear



Summary so far

- IRT modelling matches empirical data we have seen in the example ability test
- Simple models we considered so far addressed binary data (with ability applications in mind)
- There are many other applications and IRT developments in other disciplines
- Before moving on to those, need to introduce assumptions made in IRT modelling

IRT MODEL ASSUMPTIONS

IRT models

- The statistical theory is general permitting
  1. one or more traits or abilities,
  2. various model assumptions,
  3. binary or polytomous response data.
- Two IRT assumptions
  1. **dimensionality** or **local independence**
  2. **shape** of item response function (IRF)

Dimensionality or Local independence assumption

- Item responses are independent after controlling for (conditional on) the latent trait
  - or, equivalently
- There is only one dimension explaining variance in the item responses

  - The significance of these assumptions will be clear when we consider how item and person parameters are estimated

Parameter Estimation

For independent events,

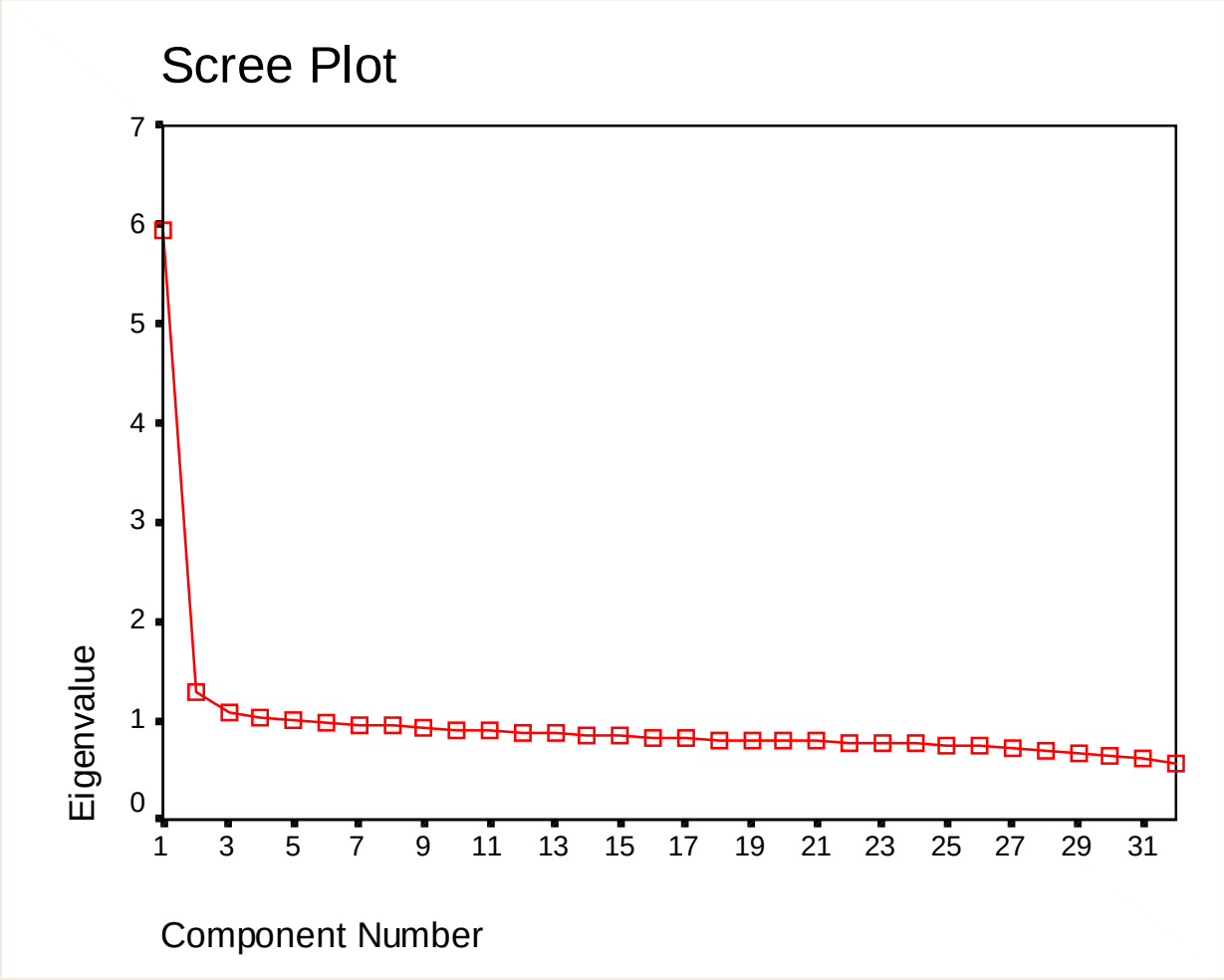
$$P(U_1,U_2,\ldots,U_n|\theta) = P(U_1|\theta)P(U_2|\theta)\ldots P(U_n|\theta) = \prod_{i=1}^{n} P(U_i|\theta)$$

When the response pattern is observed $(U_i = u_i)$

$$L(u_1,u_2,\ldots,u_p|\theta) = \prod_{i=1}^{p} P_i^{u_i} Q_i^{1-u_i}$$

where $P_i = P(u_i = 1|\theta)$ and $Q_i = 1 - P(u_i = 1|\theta)$

Checking Dimensionality Assumption:
option 1



Checking Dimensionality Assumption:
more options

- Use confirmatory approach – confirmatory item factor analysis
  - Check residuals
  - Does the unidimensional model fit?
- Cronbach's alpha is NOT an indicator of dimensionality
- Parallel analysis
  - in R package "ltm", function "unidimTest"
  - Compares empirical second eigenvalue with model-based from simulated samples

Fitting simple IRT models to binary data

PRACTICAL

Survey example

- A rural subsample of 8445 women from the Bangladesh Fertility Survey of 1989 (Huq and Cleland, 1990).
- Described in Bartholomew, D., Steel, F., Moustaki, I. and Galbraith, J. (2002) The Analysis and Interpretation of Multivariate Data for Social Scientists. London: Chapman and Hall.
- Data is available within R software package "ltm" and also on Bristol University website

The survey

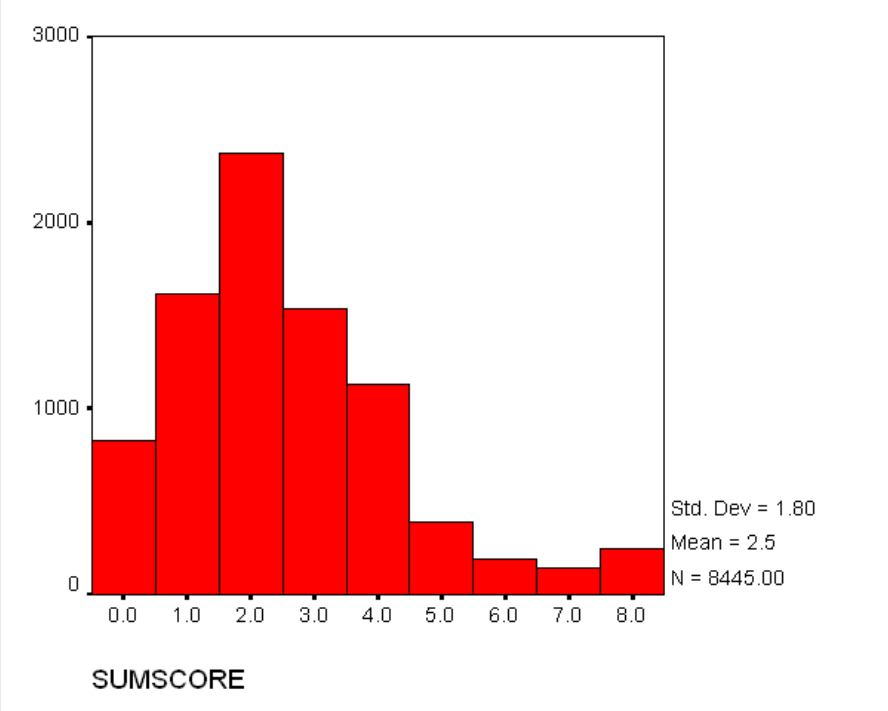
- The dimension of interest is women's mobility of social freedom.
- Women were asked whether they could engage in the following activities alone (1 = yes, 0 = no):

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.

Some frequencies

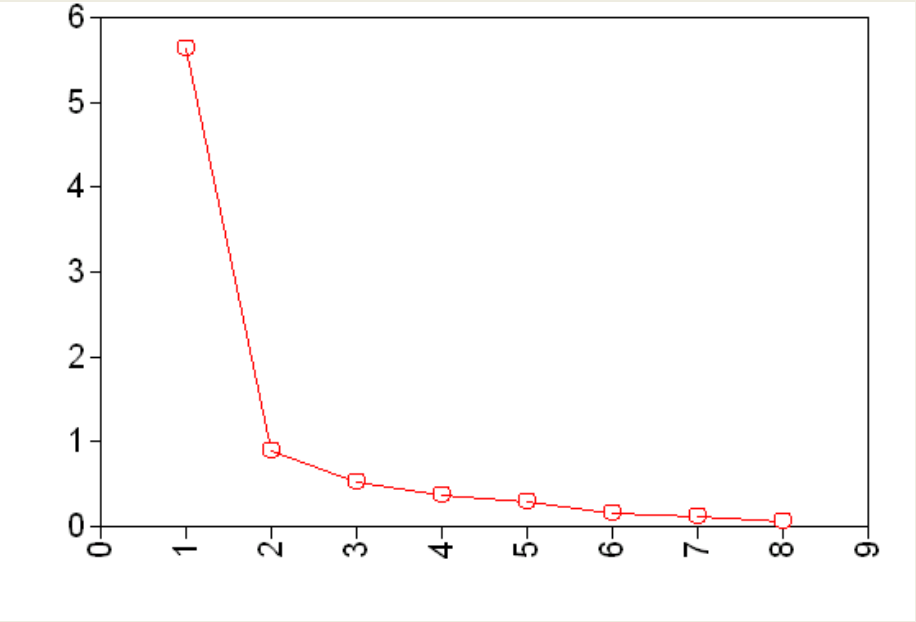
Proportions for each level of response:

	0	1	logit
Item 1	0.2013	0.7987	1.3782
Item 2	0.6861	0.3139	-0.7819
Item 3	0.2482	0.7518	1.1083
Item 4	0.6353	0.3647	-0.5550
Item 5	0.9306	0.0694	-2.5961
Item 6	0.8888	0.1112	-2.0786
Item 7	0.9470	0.0530	-2.8820
Item 8	0.9133	0.0867	-2.3549



Dimensionality

- CFA in Mplus – both full and limited information
  - Both found that 2-factor model fits significantly better
- Limited information:
  - Scree plot
  - Familiar fit indices
  - For 1 dimension
    - CFI=0.990
    - RMSEA=0.054



Dimensionality (cont.)

- Call: my2pl<-ltm(Mobility ~ z1)
  myTest<-unidimTest(my2pl)
- Output:

Unidimensionality Check using Modified Parallel Analysis

Alternative hypothesis: the second eigenvalue of the observed data is substantially larger than the second eigenvalue of data under the assumed IRT model

Second eigenvalue in the observed data: 0.8056

Average of second eigenvalues in Monte Carlo samples: 0.4889

Monte Carlo samples: 100

p-value: 0.0099

Factor loadings

- Factor loadings are relatively different

Y1	0.764
Y2	0.759
Y3	0.647
Y4	0.862
Y5	0.911
Y6	0.874
Y7	0.954
Y8	0.861

- We try to fit 2PL model

Fitting 2PL model in R

- Call: my2PL<-ltm(formula = Mobility ~ z1)
- Parameters in logistic IRT metric

DISCRIMINATION*(THETA - DIFFICULTY)

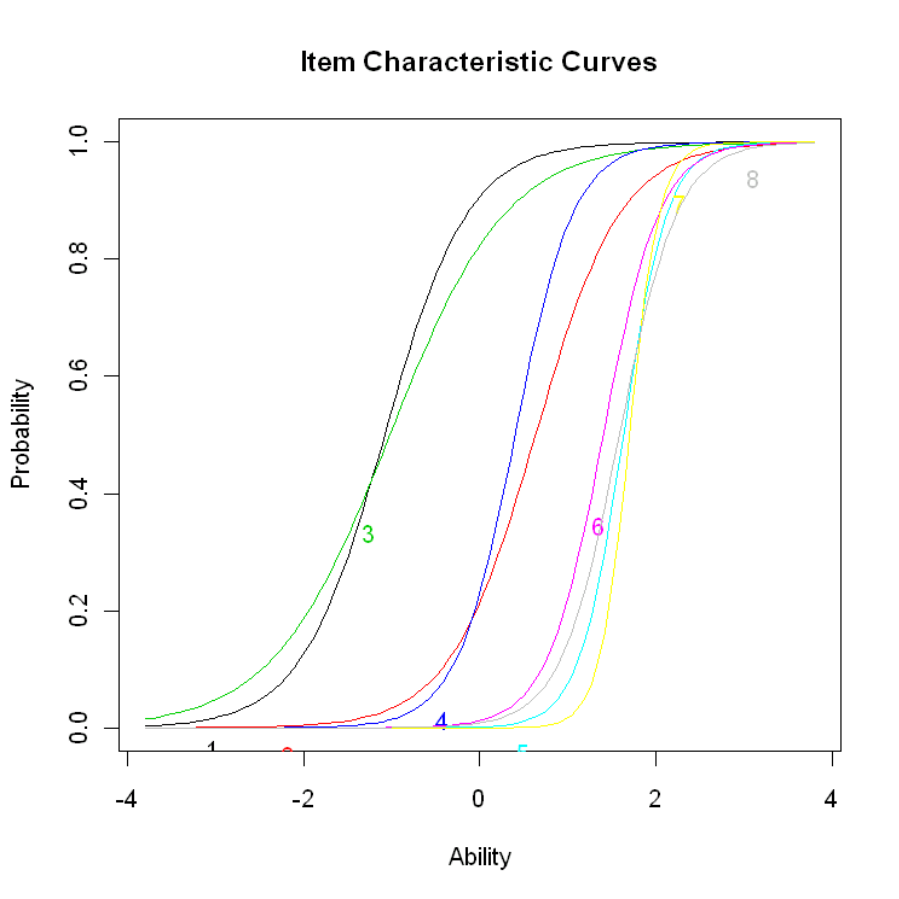
	Dffclt	Dscrmn
Item 1	-1.084	2.109
Item 2	0.631	2.058
Item 3	-1.025	1.509
Item 4	0.400	3.010
Item 5	1.630	3.976
Item 6	1.402	3.138
Item 7	1.699	5.816
Item 8	1.585	3.022

- Log.Likelihood: -23141.71

Item response functions

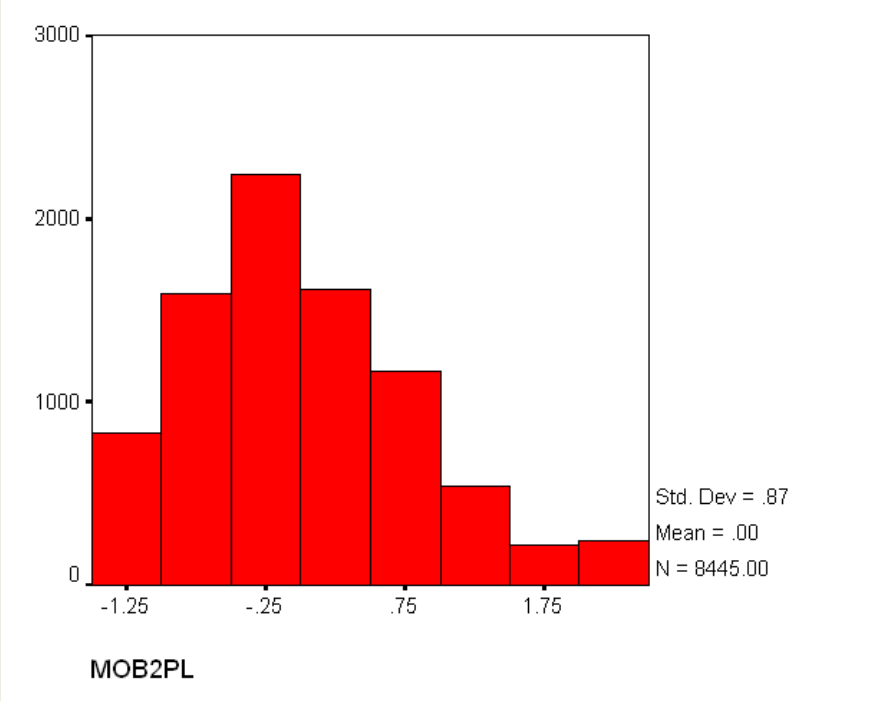
- Call:

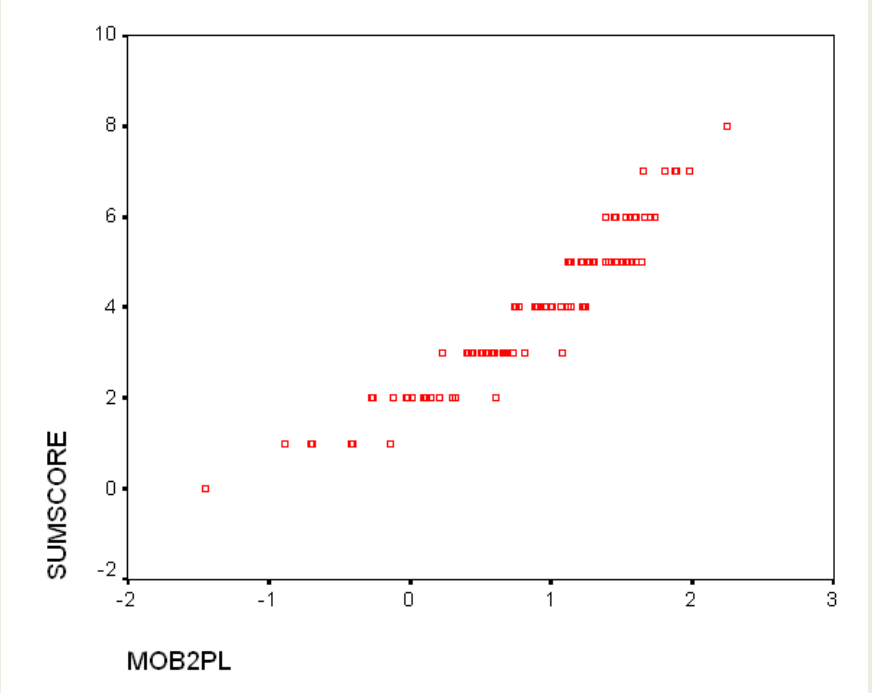
  plot(my2pl, type = "ICC")



Properties of IRT estimated scores

- Sum score and IRT estimated score correlate 0.983
- Relationship is not linear





Coming in day 2…

- More IRT models
  - More on models we introduced today
  - and new models dealing with polytomous data
- Item and test information
  - Computing SE and test reliability
- A bit about how models are estimated
- Approaches to assessing model fit