UNIVERSITY OF CAMBRIDGE

The Psychometrics Centre

Summer School in Applied Psychometric Principles

Peterhouse College

13th to 17th September 2010

Two- and three-parameter IRT models. Introducing models for polytomous data. Test information in IRT and reliability. Testing assumptions and assessing model fit.

Day 2

Anna Brown, PhD
University of Cambridge

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Topics covered yesterday

- We have…
  - Introduced IRT
  - Introduced simple models for binary responses
  - Mentioned the main IRT assumptions
  - Tested 2PL model with Mobility survey data

Topics to cover today

- Item and examinee parameter estimation
- IRT models and their properties
  - IRT models for binary data (more formal treatment)
  - IRT models for polytomous data (questionnaires and surveys with multiple answer options, essays etc.)
- Item and test information; reliability in IRT
- Assessing model fit
- Summary – selecting an appropriate IRT model

How item parameters and examinee scores are estimated

ITEM AND EXAMINEE PARAMETER ESTIMATION

Likelihood of item responses

For independent events,

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

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

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

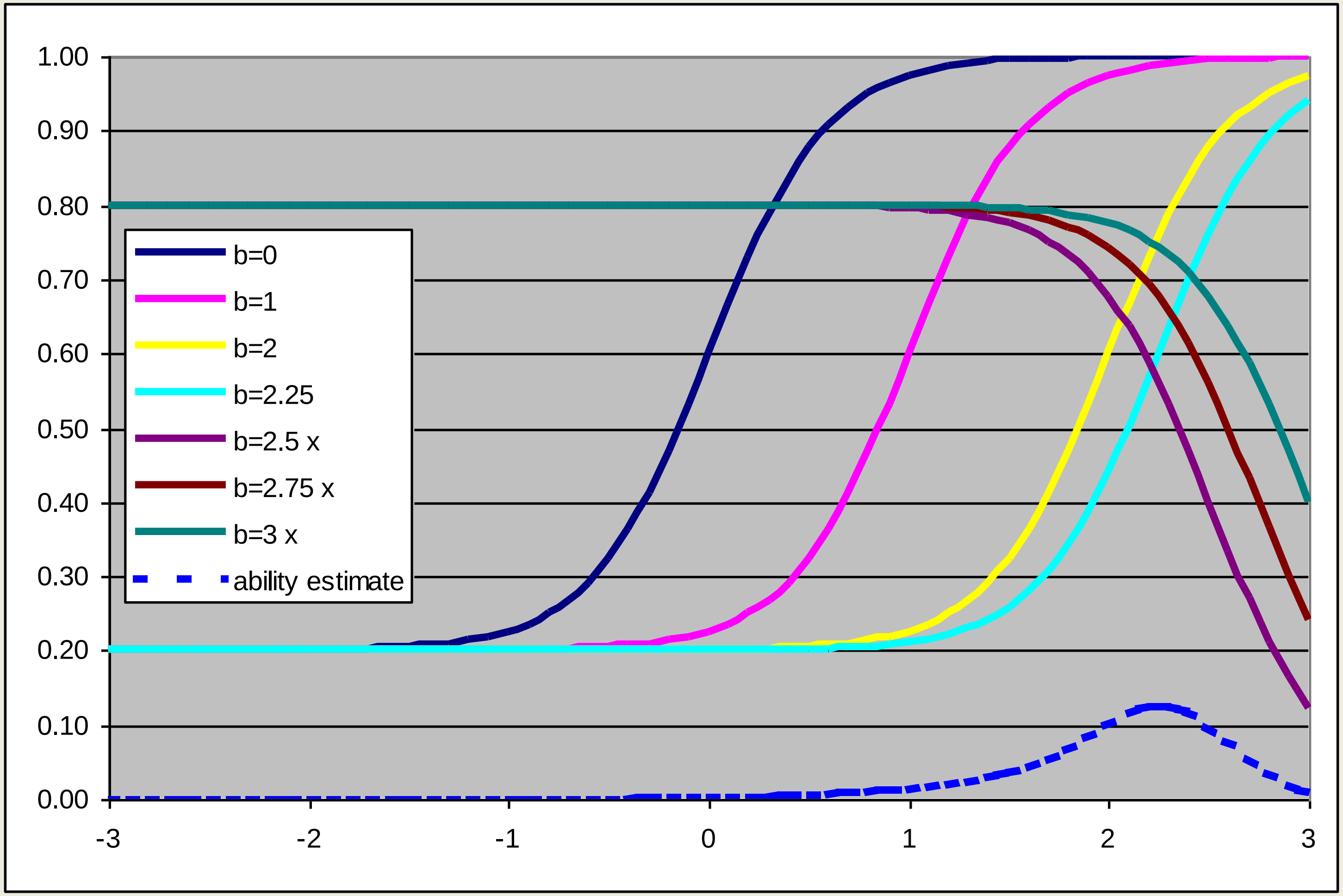
where $P_i = P(u_i = 1|\theta)$ and $Q_i = 1 - P(u_i = 1|\theta)$

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Estimating examinee parameters

- In routine applications of tests item parameters will be known (calibrated during standardisation)
- Given individual pattern of item responses, probabilities of responses will depend only on the latent trait
- Assuming responses are independent after controlling for the latent trait, the joint probability of the response pattern *equals* the product of probabilities of responses to individual items

Probabilities of responses to several items



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Finding the examinee parameter

- Maximum likelihood (ML)
  - Maximising the likelihood function (iterative process)
  - ML estimator is unbiased, and its errors are normally distributed
  - Problems with ML is that convergence is not guaranteed with aberrant responses, and no estimator exists for all correct/incorrect responses
- Maximum a posteriori (MAP)
  - Maximises the mode of the posterior distribution (iterative process); implemented in Mplus
  - Estimator exists for all response patterns, more precise
  - Biased towards the sample mean
- Expected a posteriori (EAP)
  - Maximises the mean of the posterior distribution (non-iterative)
  - Estimator exists for all response patterns, more precise
  - Biased towards the sample mean

Estimating item parameters

- Joint maximum likelihood estimation (JML)
  - Uses *observed* frequencies of response patterns
  - Starting values for ability as proportion correct
    1. Estimate item parameters
    2. Use item parameters to re-estimate ability
  - Repeat last two steps until estimates do not change
- Marginal maximum likelihood (MML)
  - Uses *expected* frequencies of each response pattern
  - EM (Estimation and Maximisation) by Bock & Aitken (1981) is popular
- Conditional maximum likelihood (CML)
  - Uses sufficient statistics to exclude trait level parameters (only applies to the Rasch models)

Estimation issues

- Test assumptions
  - Unidimensionality or Local independence
    - Unspeeded data in ability tests
- Model fit
- Data requirements (only guidelines)
  - 1 parameter – n>200
  - 2 parameter – n>600
  - 3 parameter – n>1000

Options for binary and polytomous data

IRT MODELS FOR YOUR DATA

IRT modelling options

Outcome	IRT models
Binary	Binary IRT (1PL (Rasch), 2PL, 3PL)
Polytomous	
Nominal	Nominal response model (2PL)
Ordinal	Graded Response family (2PL), Partial Credit family (2PL)

Over 100 IRT models in the testing field, but really only 8 to 10 in wide use (van der Linden & Hambleton, 1997).

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

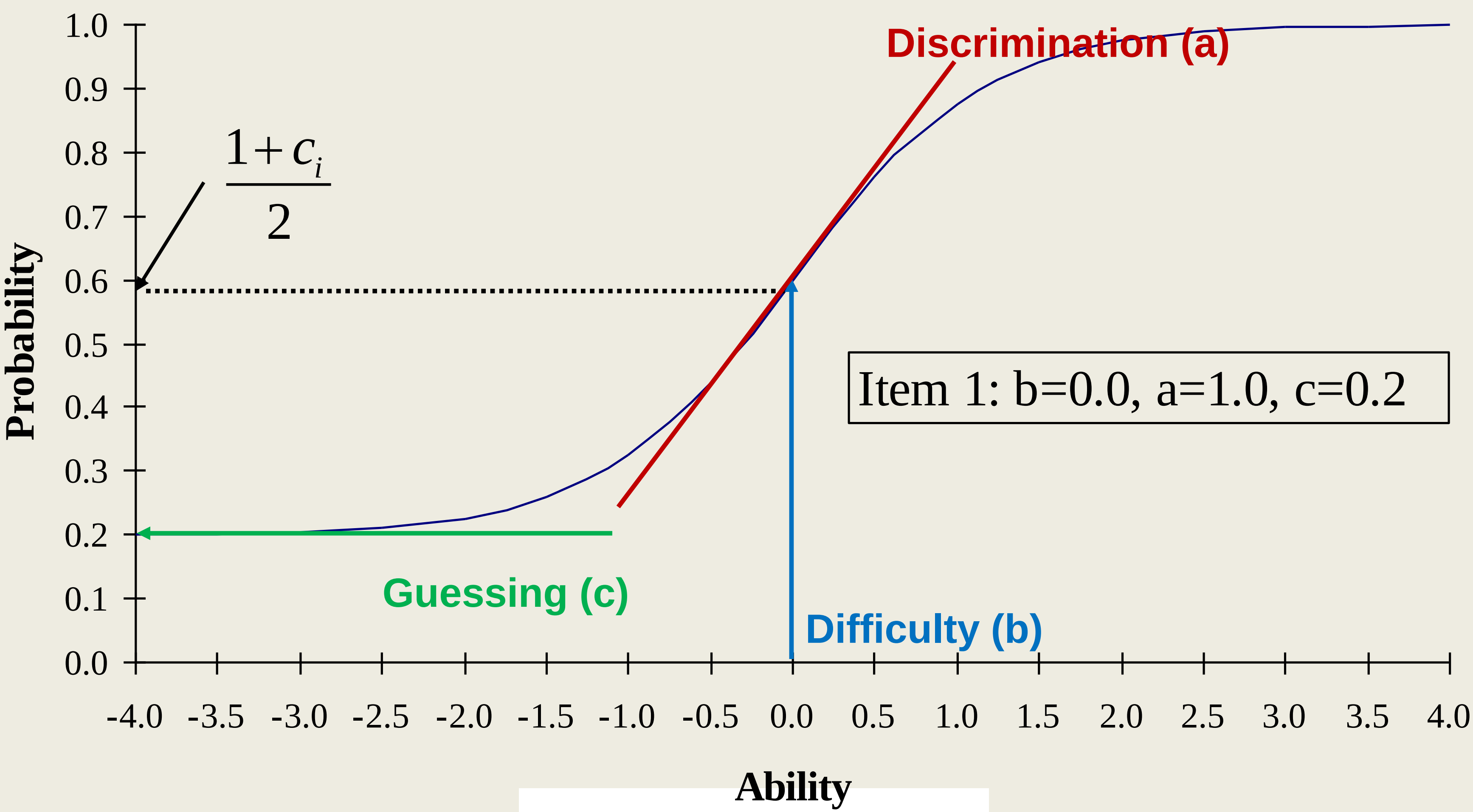
Three-Parameter Logistic Model:

- This model is suitable for item responses to multiple choice items scored correct/incorrect

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

  - In speeded tests and exams, probability of success even for difficult items might never fall below certain level
  - Guessing parameter is typically close to 1 divided by the number of alternatives

Item parameters for the 3PL model



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Two-Parameter Logistic Model:

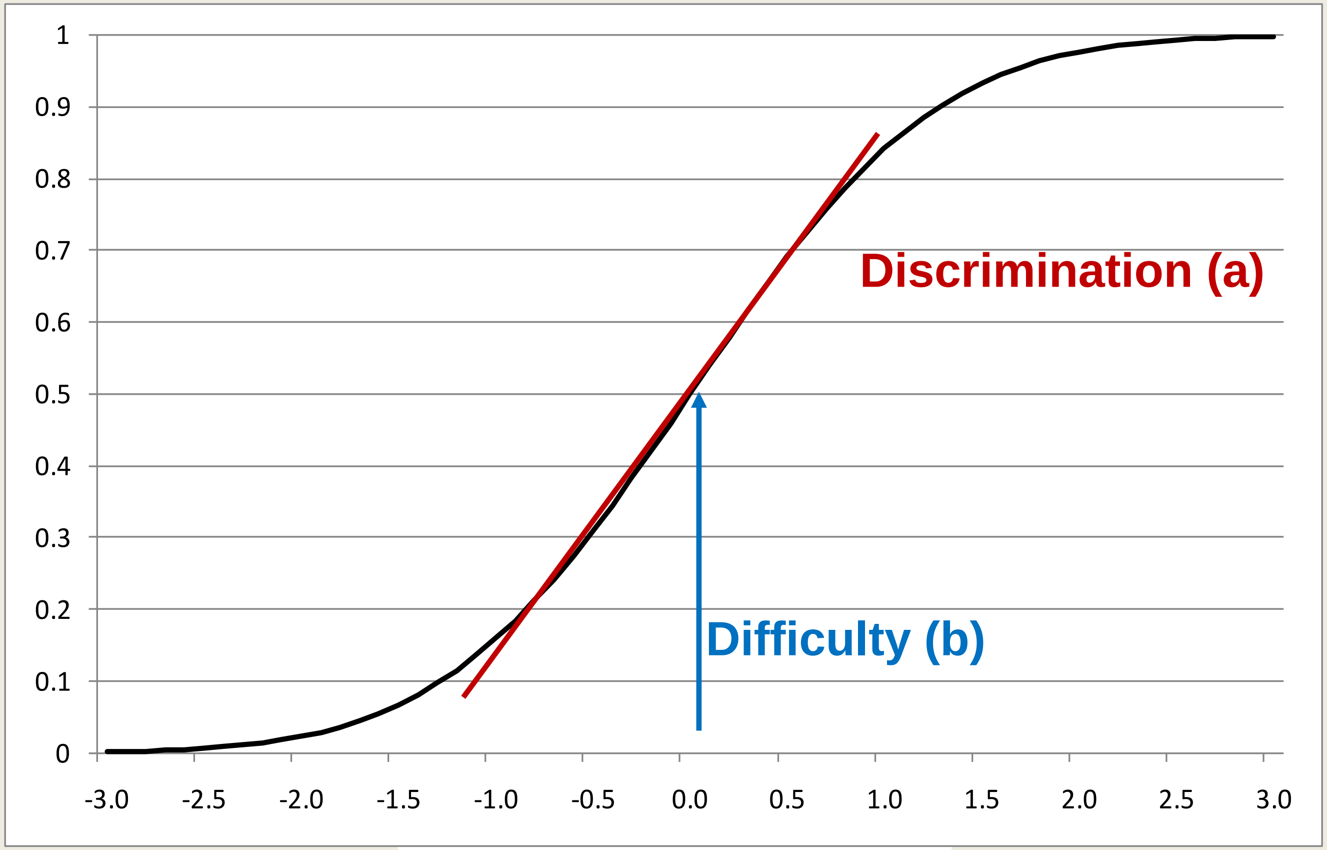
- This model is suitable for many types of binary item responses

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

  - To ability items scored correct/incorrect (without guessing)
  - To “yes/no” “agree/disagree” type responses to questionnaire items
  - Accommodates different factor loadings and negatively keyed items

Item parameters for the 2PL model

- Parameters: a=1, b=0



Interpretation of Item Parameters for the Logistic Models

- Reporting scale is only defined up to a linear transformation $b^* = xb + y$
- Common to set ability scores to a mean of 0.0 and a standard deviation of 1.0
  - In Rasch model, average b value is often set to zero instead
- An assumption of ability being normally distributed does NOT need to be made
- On this scale (with D=1.7 in the model), b values [-2.0, +2.0], a values [0.0, 2.0], and c values [00, .25] are common

Practical (Ability.dat)

- Let's fit 2PL and 3PL models to 20-item ability test data in R

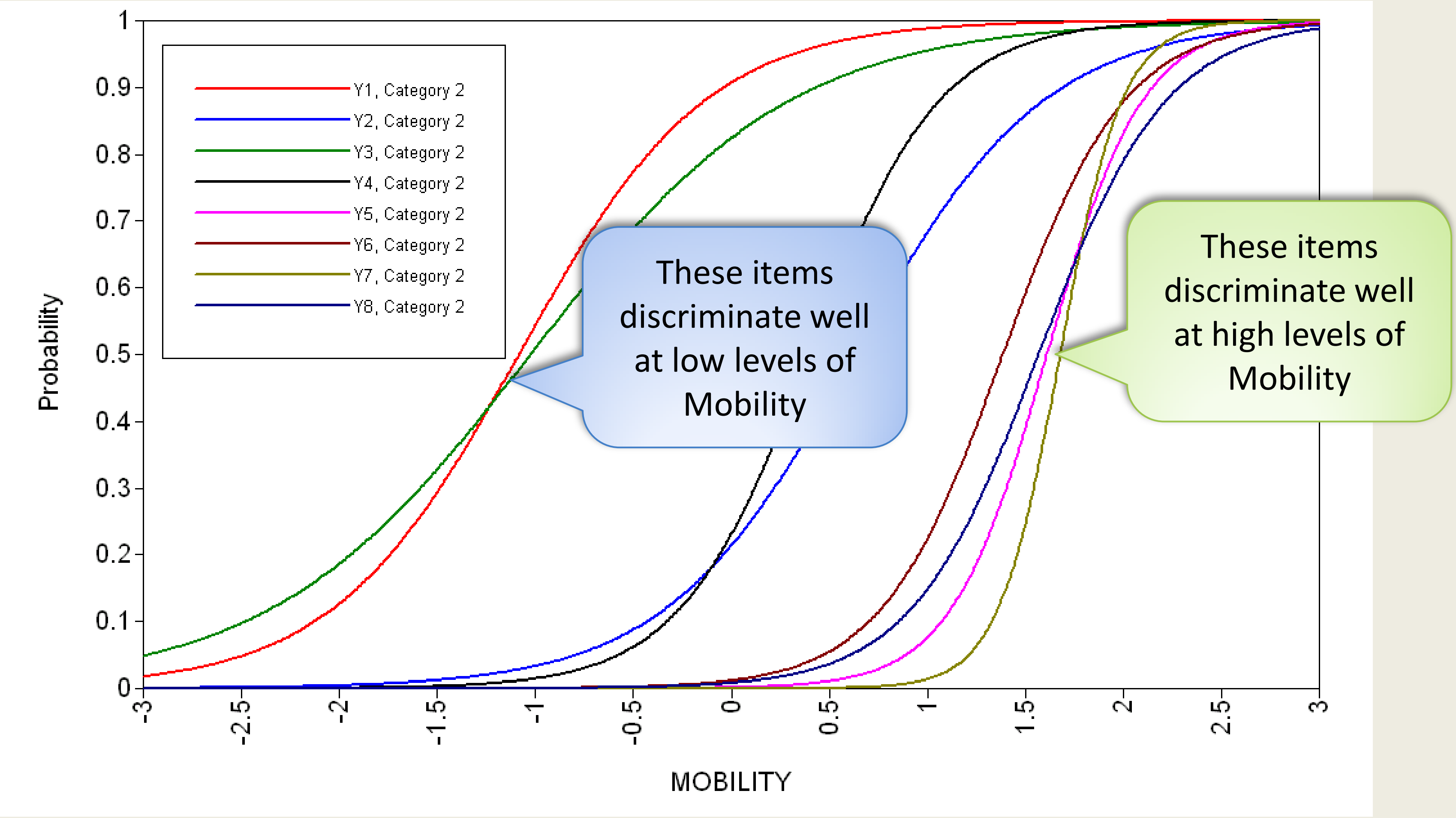
Test reliability in Item Response Theory

INFORMATION AND MEASUREMENT ERROR

Reliability in IRT

- Items may have different discrimination power
- Items discriminate better around their difficulty parameter
  - An easy item is useless at discriminating between examinees of high ability (they all will get it right)
  - A difficult item is useless at discriminating between examinees of low ability (they all will get it wrong)
- In contrast with CTT, in IRT reliability varies for different levels of the latent trait

IRFs for our Mobility survey



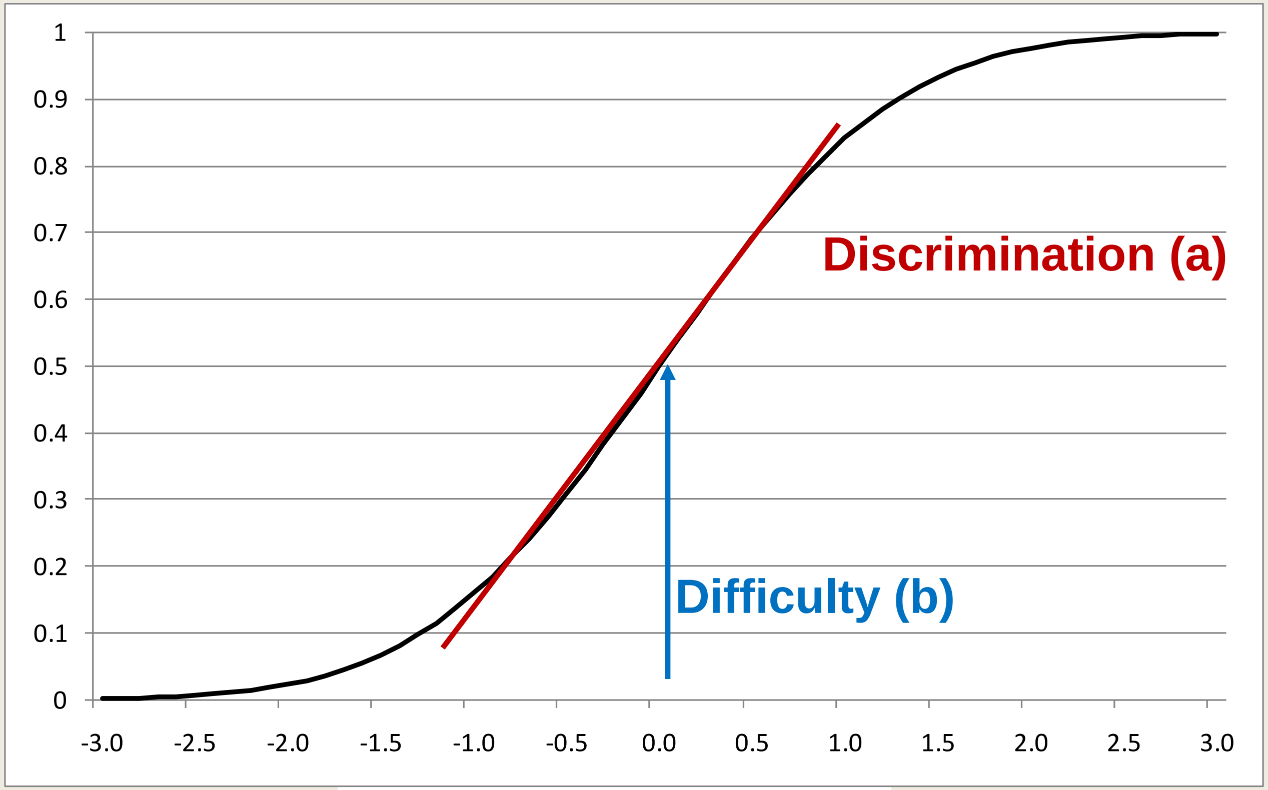
UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Item information

- The concept of the gradient of a function z=f(x)
  - change in z corresponding to a an increase in x
  - slope of a local tangent to the curve at each point
  - item discrimination parameter in 2PL model reflects the slope of a tangent at the curve inflection point (item difficulty)
- Derivative f'(x) is a relative change in f(x) when x increases by an infinitely small amount

Example IRF

- With parameters a=1, b=0



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Item Information Function (IIF):

$$I_i(\theta) = \frac{\left[P'_i(\theta)\right]^2}{P_i(\theta)\left[1 - P_i(\theta)\right]}$$

- The amount of information the item provides about the latent trait
- Analytical expressions for derivatives of both logistic and normal-ogive functions are easy to derive
- Then they can be substituted in the formula

IIFs for logistic models

- For 3PL model (remember constant D=1.7?)

$$I_i(\theta) = \left[1.7a_i(1-c_i)\right]^2 P_i(\theta)\left[1-P_i(\theta)\right]$$

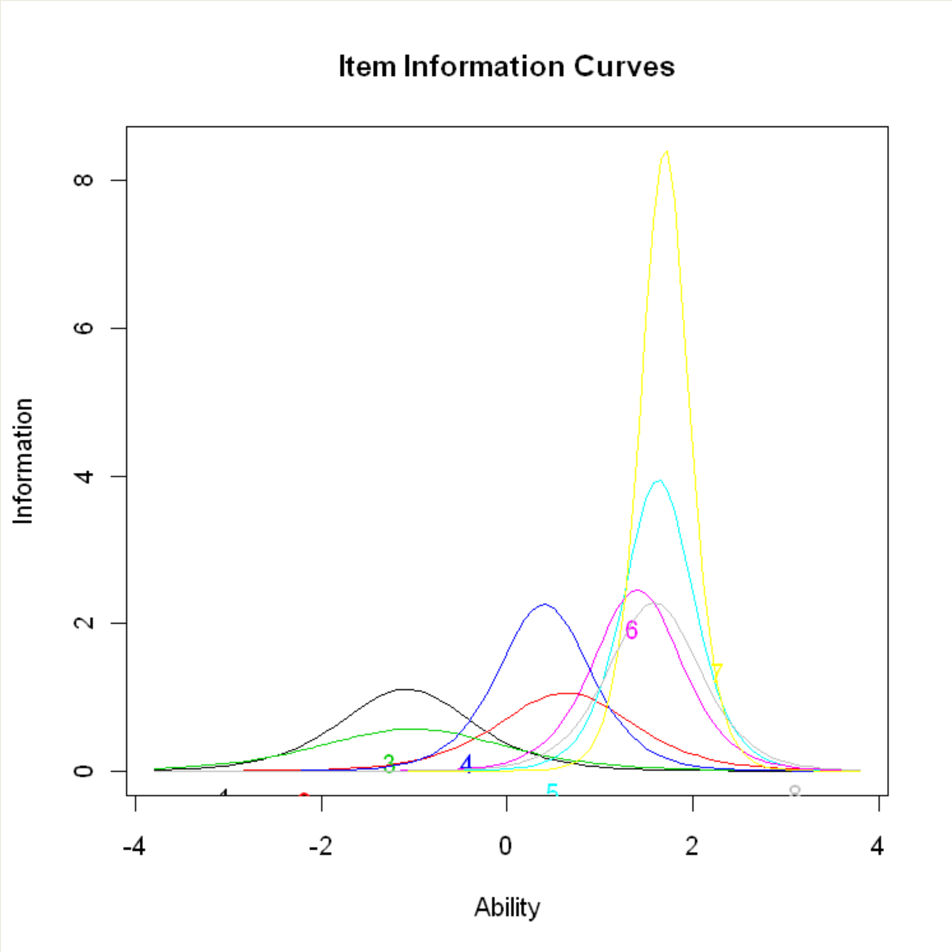
- For 2PL model

$$I_i(\theta) = \left[1.7a_i\right]^2 P_i(\theta)\left[1-P_i(\theta)\right]$$

- For 1PL model (discrimination is constant)

$$I_i(\theta) = \left[1.7a\right]^2 P_i(\theta)\left[1-P_i(\theta)\right]$$

IIFs for the Mobility survey

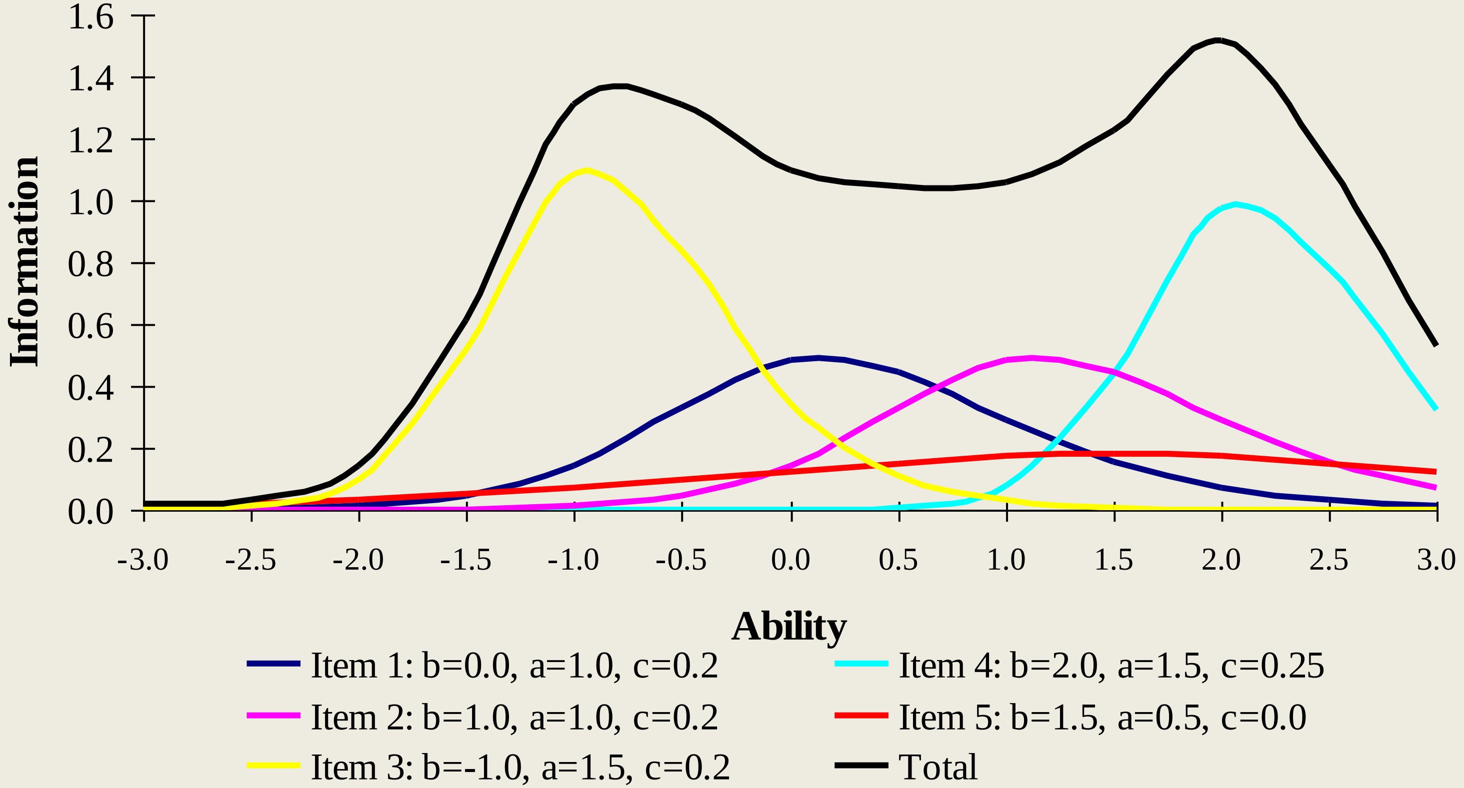


Test information

- Test information is the sum of all item information functions
  - Providing that the local independence holds

$$I\left(\theta\right)=\sum_{i=1}^{p} I_i\left(\theta\right)$$

IIFs and TIF



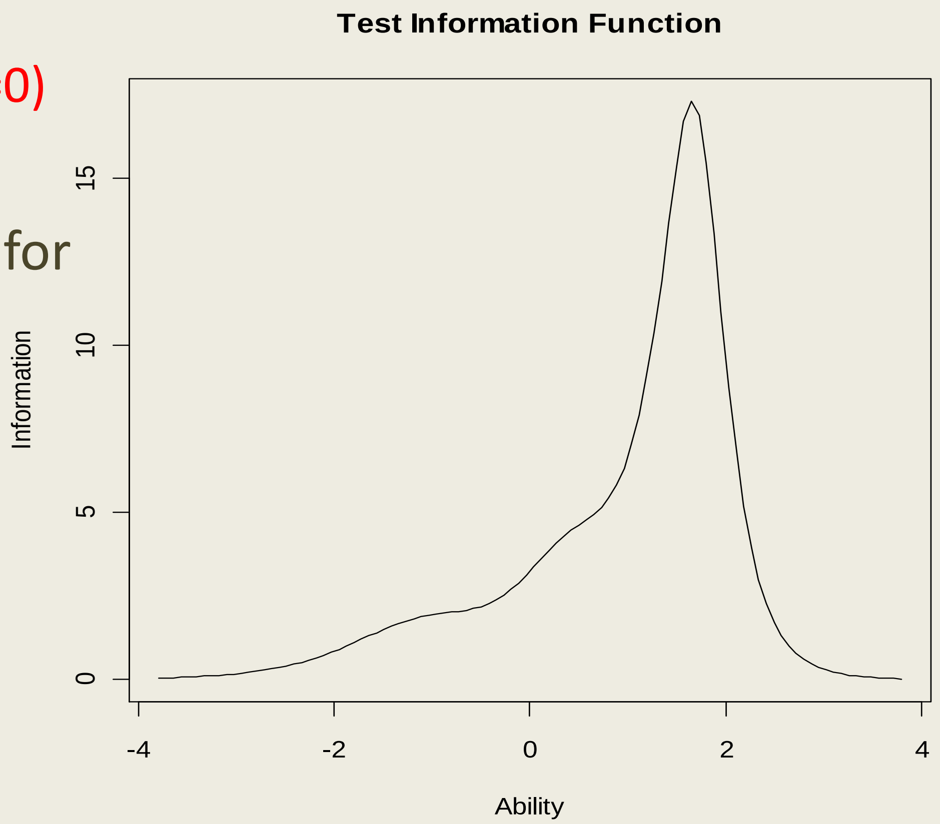
UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

TIF for the Mobility survey

- To obtain test information in R:

  > plot(my2pl, type = "IIC", items=0)

- In Mplus, information is scaled for the logistic model (with 1.7 scaling constant)
- If using normal ogive model (which is the default in Mplus), multiply given values by 2.89 (1.7^2).



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

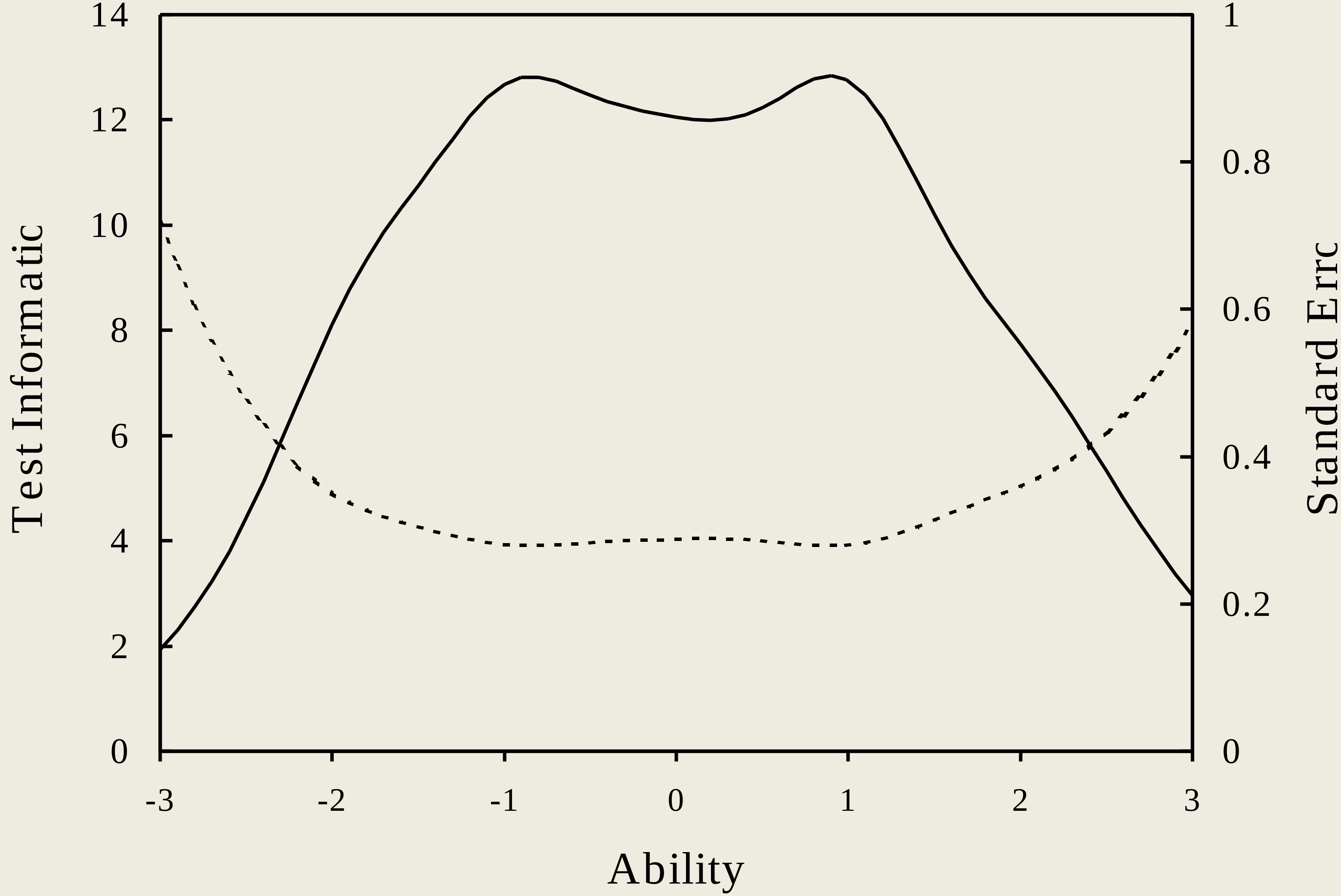
Information and Standard Error

- Error of measurement inversely related to information
- Standard error (SE) is an estimate of measurement precision at a given theta
- SE = inverse of the square root of the item information

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

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

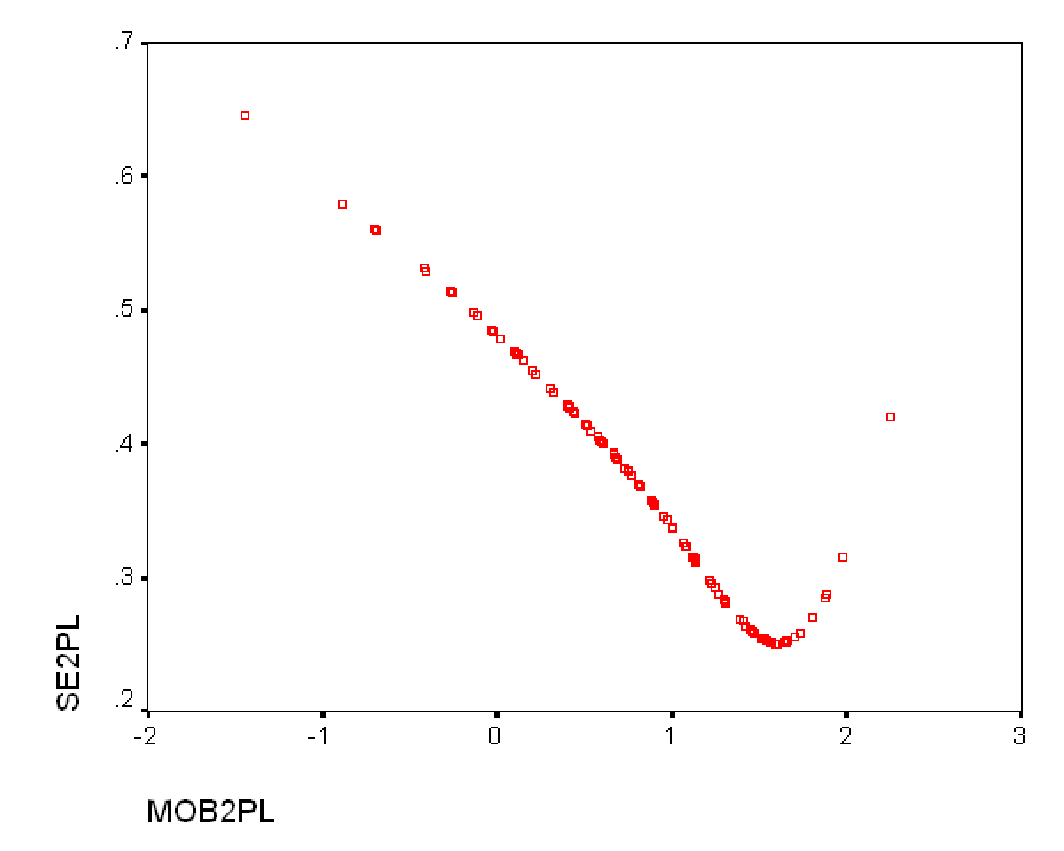
TIF & Standard Errors



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Mobility data Standard Errors

- Plotting empirical SEs for each individual



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Reliability in IRT

- Test reliability in CTT is defined as the proportion of variance in the test scores due to the true score
- This can easily be extended to IRT
  - True score is the latent trait
  - Score variance is the sum of the latent trait variance and the error variance
  - Error variance σ_e is the squared SE, or reciprocal of test information

.

$$\sigma^2_{error}(\theta) = SE^2(\theta) = \frac{1}{I(\theta)}$$

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Practical (Ability.dat)

- Obtain and assess Item Information curves to 20-item ability test data in R
- Obtain and assess Test Information curves
- Can we estimate the test reliability?

Theoretical and empirical IRT reliabilities

- Single index of reliability might be desirable in applications
  - Error variance must be summarised across the latent trait (when the information is relatively uniform)
- IRT theoretical reliability
  - Assume trait variance is 1
  - Squared SEs are averaged across the latent trait (integration is required)

$$\rho_t = 1 - \bar{\sigma}^2_{error}$$

- IRT empirical reliability
  - True variance = observed minus error
  - Squared SEs are averaged across estimated values in the sample

$$\rho_e = 1 - \frac{\bar{\sigma}^2_{error}}{\sigma^2}$$

POLYTOMOUS RESPONSE MODELS

Polytomous Response Models

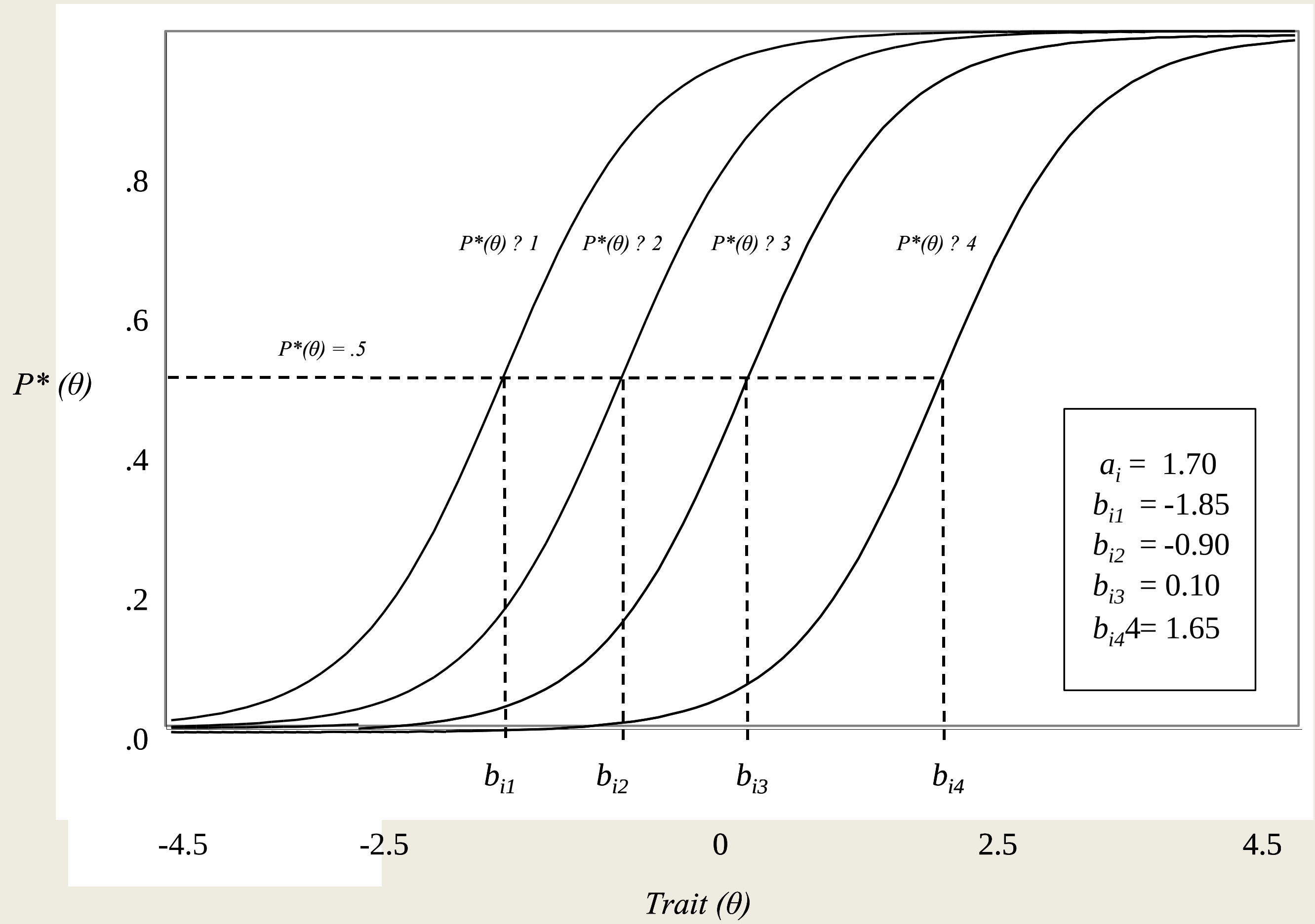
- Responses to items might be in more than two categories
- Models to handle essay scores, Likert scales, other rating scales, etc.
  - Graded Response Model (Samejima, 1969; 1996) and its variations
  - Partial Credit Model (Masters, 1982) and its more general version (Muraki, 1992)
  - Nominal Response Model (Bock, 1972)

GRADED RESPONSE MODELS

The Graded Response logic

- Extension of the 2PL model to handle multiple response categories that are logically ordered
- Computing probability of response to each category requires a 2-step process:
  - First, probability of responding in or above category x, P_x*, is computed
    - These are simple 2PL curves reflecting the dichotomy
  - Second, probability of responding in category x equals the difference $P_x* - P_{x+1}*$

Cumulative score category functions for a 5-category item



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

The Graded Response Model

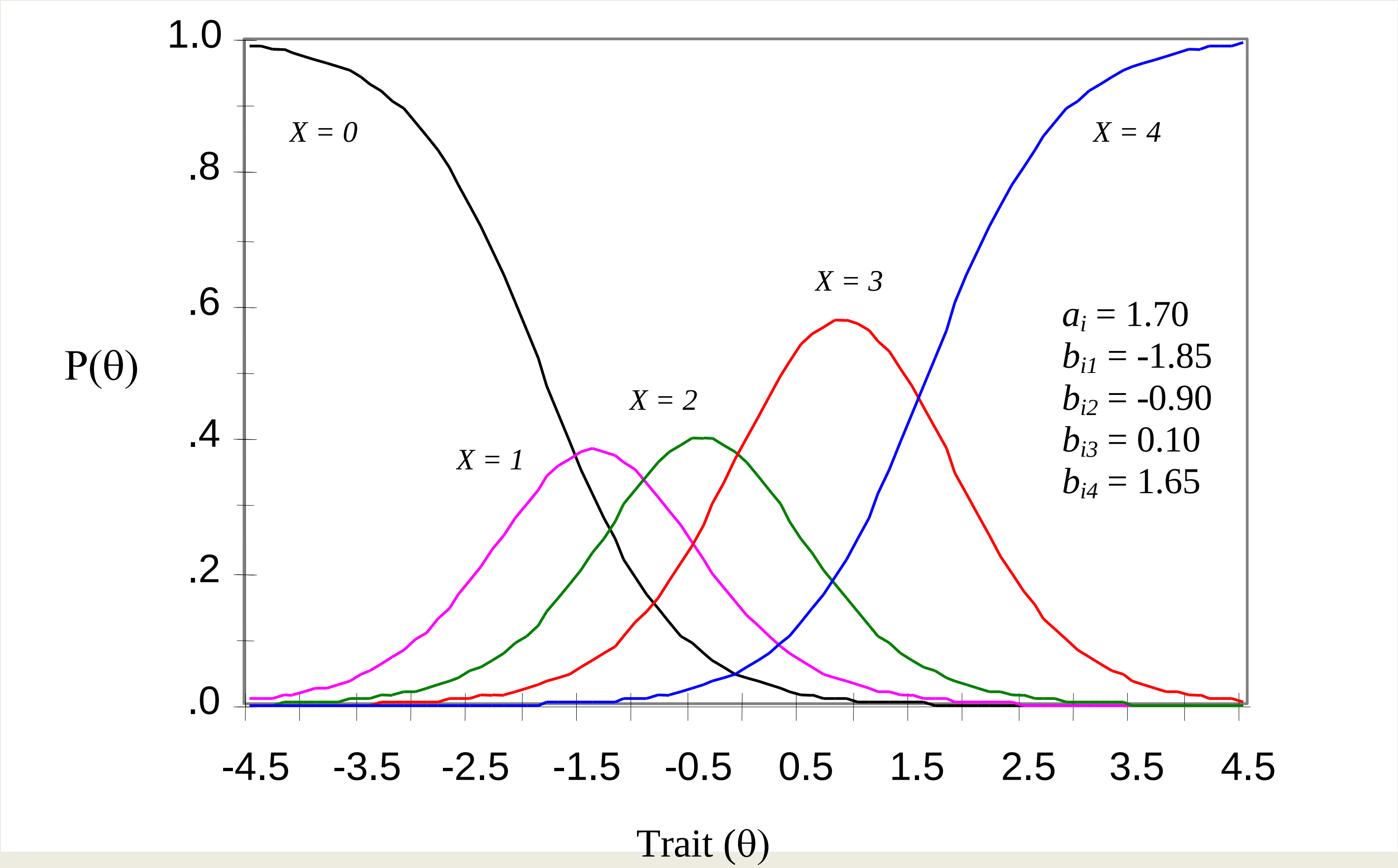
- Let $x = 0,1,\ldots, m_i$ be a category number
  - the number of categories can vary between items!
- Then
  - probability of responding in the lowest category or above is 1 ($P^*_0=1$)
  - Probability of responding in the highest category is $P_{mi}= P^*_{mi}$
  - Probability of responding in any intermediate category is $P_x= P^*_{mx}- P^*_{mx+1}$
- Probability of falling in the category x or above is

$$P^*_{ix}(\theta) = \frac{e^{Da_i(\theta - b_{ix})}}{1 + e^{Da_i(\theta - b_{ix})}}$$

- Item has one discrimination (a_i) and m_i threshold parameters (b_{ix})

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Score category functions for a 5-category item



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Features of the GRM

- Very widely applicable to questionnaire data
  - Items can have different discriminations
  - Items can have different number of categories
    - Do not have to worry about 0 responses in a particular category
  - Category thresholds can be spaced at any intervals (and this is extremely flexible compared to the equidistant coding assumption of the Likert scale)
    - Do not have to worry about whether distance between "never" and "rarely" is the same as between "sometimes" and "often"
  - Category thresholds have to be ordered – a very reasonable assumption in most questionnaires using rating scales

The Modified GRM

- Muraki (1990) developed a model suitable for items using the same rating scale
- Restricted version of GRM, where
  - Slopes (a_i) vary between items
  - Threshold parameters are partitioned into two terms:
    - One location parameter (b_i) for each item *i*
    - m category threshold parameters ($c_1 \ldots c_m$) for the entire scale
- "Restricted" because assumes that category boundaries are equally distant across items
  - Has fewer parameters
  - Scale for parameters c is arbitrary

Practical (Big5.dat)

- Big Five personality factors (Goldberg, 1992)
  - Extraversion (or Surgency), Agreeableness, Emotional stability, Conscientiousness and Intellect (or Imagination)
- IPIP (International Personality Item Pool), 60-item questionnaire measuring the Big Five
  - 12 items per trait
  - 5 symmetrical rating options:

    Very Inaccurate / Moderately Inaccurate / Neither Accurate Nor Inaccurate / Moderately Accurate / Very Accurate
- Volunteer sample, N=438 (52% female, 48% male)

– Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4, 26-42.*

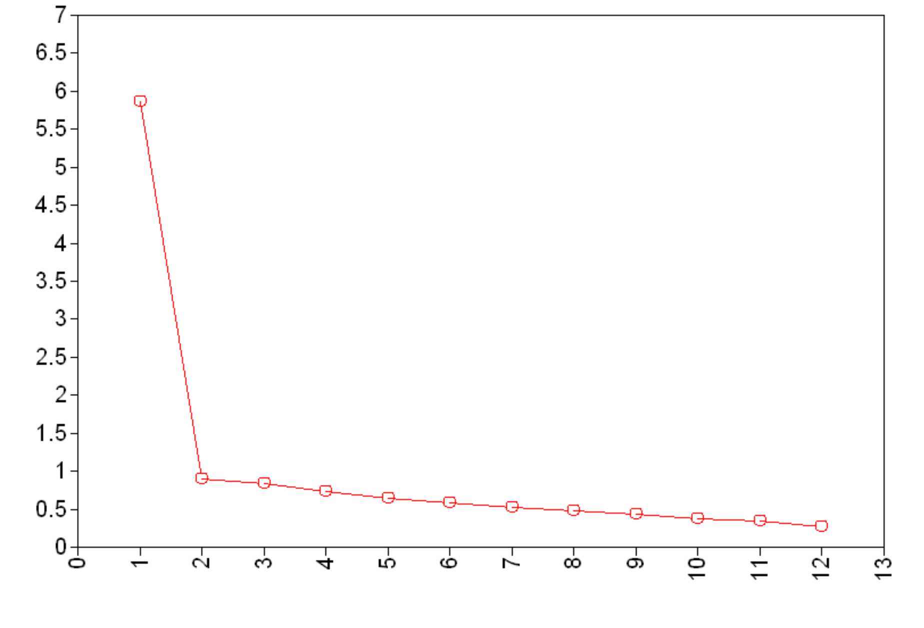
Extraversion

- 12 items, 8 positive and 4 negative

No	Item	Key
13	I start conversations	1
14	I am the life of the party	1
15	I feel at ease with people	1
16	I am quiet around strangers	-1
17	I keep in the background	-1
18	I don't talk a lot	-1
19	I talk to a lot of different people at parties	1
20	I feel comfortable around people	1
21	I find it difficult to approach others	-1
22	I make friends easily	1
23	I don't mind being the centre of attention	1
24	I am skilled in handling social situations	1

Checking assumptions

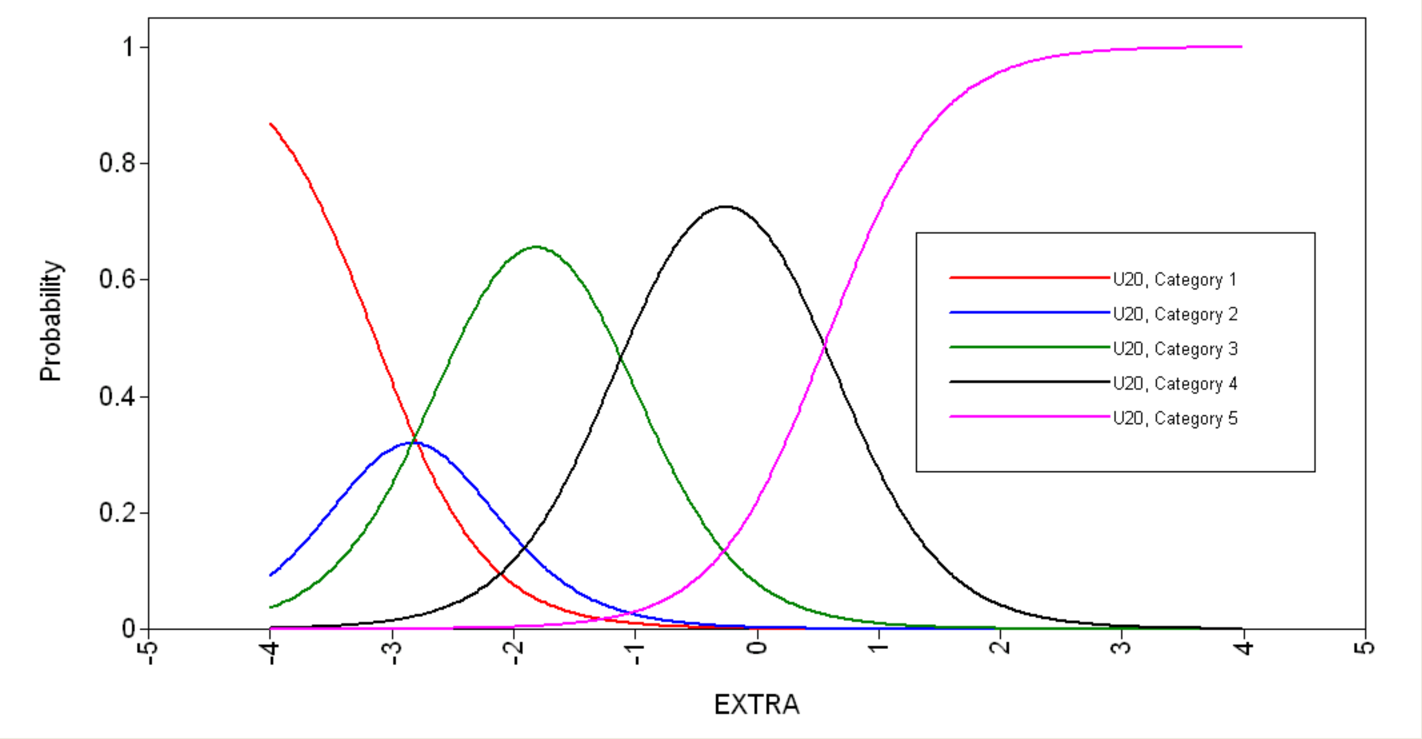
- CFA in Mplus
  - Chi-square 218.681 (df=54); CFI=0.959; RMSEA=0.083
- Essentially unidimensional



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

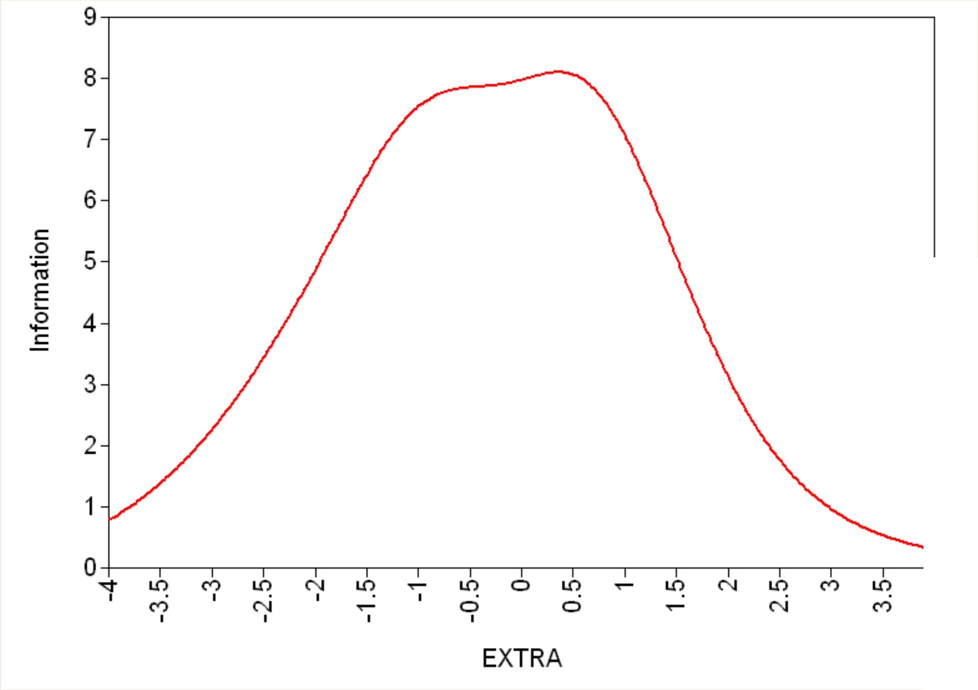
IRFs for item 20

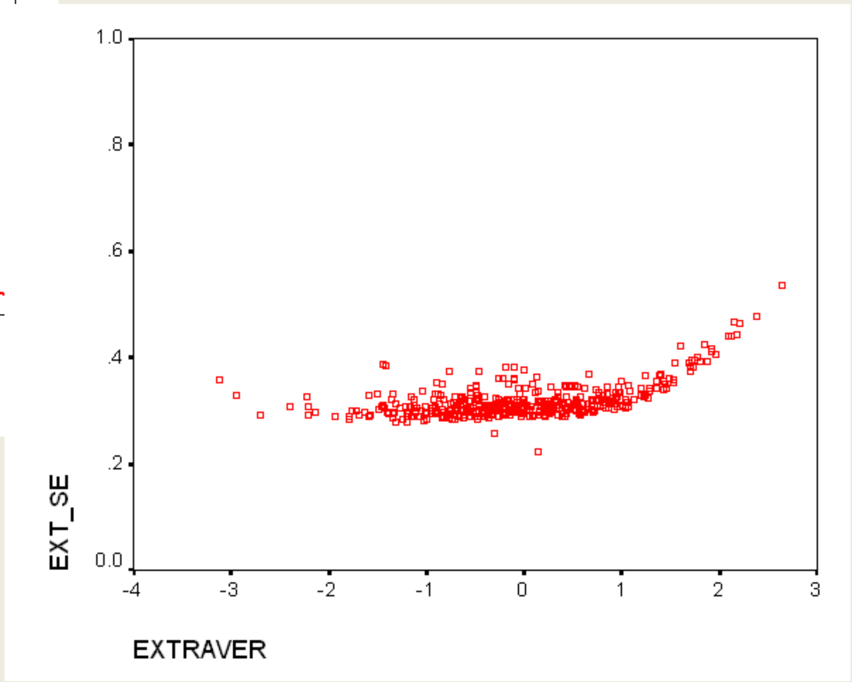
- **“I feel comfortable around people”**
- Highest discrimination parameter (a=2.19)



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Test information and SEs





SEs and reliability for the sample

- Mplus now outputs SEs of the estimated trait score
- Empirical reliability can easily be computed

$$\rho_t = \frac{\sigma^2 - \bar{\sigma}^2_{error}}{\sigma^2}$$

  - Ave squared SE = 0.114
  - Observed variance = 0.899
  - Empirical reliability is (0.899-0.114)/0.899=0.87

PARTIAL CREDIT MODELS

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

The Partial Credit logic

- Created specifically to handle items that require logical steps, and partial credit can be assigned for completing some steps (common in mathematical problems)
- Completing a step assumes completing **all steps** below
- Computing probability of response to each category is direct ("divide-by-total"):
  - Probability of responding in category x (completing x steps) is associated with ratio of
    - odds of completing all steps before and including this one, and
    - odds of completing all steps
  - Each step's odds are modelled like in binary logistic models
    - For an item with m+1 response categories, m ***step difficulty*** parameters $b_1 \ldots b_m$ are modelled

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Generalized Partial Credit Model

- The model is:

$$P_{ix}(\theta) = \frac{\exp \sum_{s=0}^{x} a_i(\theta - b_{is})}{\sum_{r=0}^{m}\left[\exp \sum_{s=0}^{r} a_i(\theta - b_{is})\right]}$$

- Easier to see step by step (assume 3 categories):
  - Probability of completing 0 steps

$$P_{i0}(\theta) = \frac{\exp[0]}{\exp[0] + \exp\left[0 + a_i(\theta - b_{i1})\right] + \exp\left[0 + a_i(\theta - b_{i1}) + a_i(\theta - b_{i2})\right]}$$

  - Probability of completing 1 step

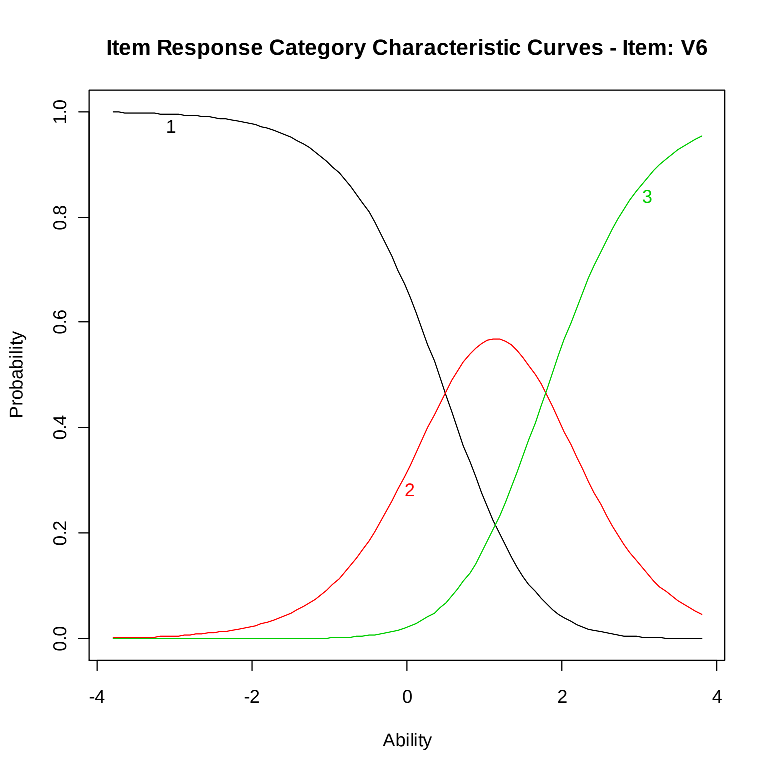
$$P_{i0}(\theta) = \frac{\exp\left[0 + a_i(\theta - b_{i1})\right]}{\exp[0] + \exp\left[a_i(\theta - b_{i1})\right] + \exp\left[0 + a_i(\theta - b_{i1}) + a_i(\theta - b_{i2})\right]}$$

  - Etc. .. Easy to see that it is “divide-by-total” model, which for 2 categories reduces to 2PL model

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Item response functions for GPCM

- Step difficulty parameters have an easy graphical interpretation – they are points where the category lines cross
- Relative step difficulty reflects how easy it is to make transition from one step to another
  - Step difficulties do not have to be ordered
  - "Reversal" happens if a category has lower probability than any other at all levels of the latent trait
- Lines nicely reflect how frequently each category is selected



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Applications of GPCM

- Cognitive tasks where giving credit for partial completion are the obvious applications
- Used often for rating scales as well
  - (though it is less clear how the logic of partial credit applies to some of them)
  - Research shows that GRM and GPCM applied to the same polytomous questionnaire data produce virtually identical results

Practical (SDQ_R.dat)

- Strengths and Difficulties Questionnaire (Goodman, 1997)
- Emotional symptoms subscale (5 items)
    1. I get a lot of headaches, stomach-aches or sickness
    2. I worry a lot
    3. I am often unhappy, down-hearted or tearful
    4. I am nervous in new situations. I easily lose confidence
    5. I have many fears, I am easily scared
- Response categories

  not true – somewhat true – certainly true

NOMINAL RESPONSE MODELS

Nominal responses

- What about items where ordering of categories does not make sense or is not obvious?
  - Distracter alternatives in multiple choice cognitive items
    - Of course simple correct/incorrect scoring will do in most cases but some distracters can be "more correct than others" and therefore provide useful information
  - Questionnaire items with response options that are not rating scale (e.g. possible alternatives for attitudes or behaviours)
    - In a measure of risk for bulimia: "***I prefer to eat***"

(a) at home alone - (b) at home with others – (c) in a restaurant – (d) at a friend's house – (e) doesn't matter

Nominal response model

- Bock (1972) proposed another "divide-by-total" model

$$P_{ix}(\theta) = \frac{\exp(a_{ix}\theta - c_{ix})}{\sum_{x=0}^{m} \exp(a_{ix}\theta - c_{ix})}$$

- Notice that:
  - Each category has its own discrimination parameter a_x (and these can be positive and negative)
  - Each category has its own intercept parameter c_x
  - To identify the model, constraints on a_x and c_x must be set

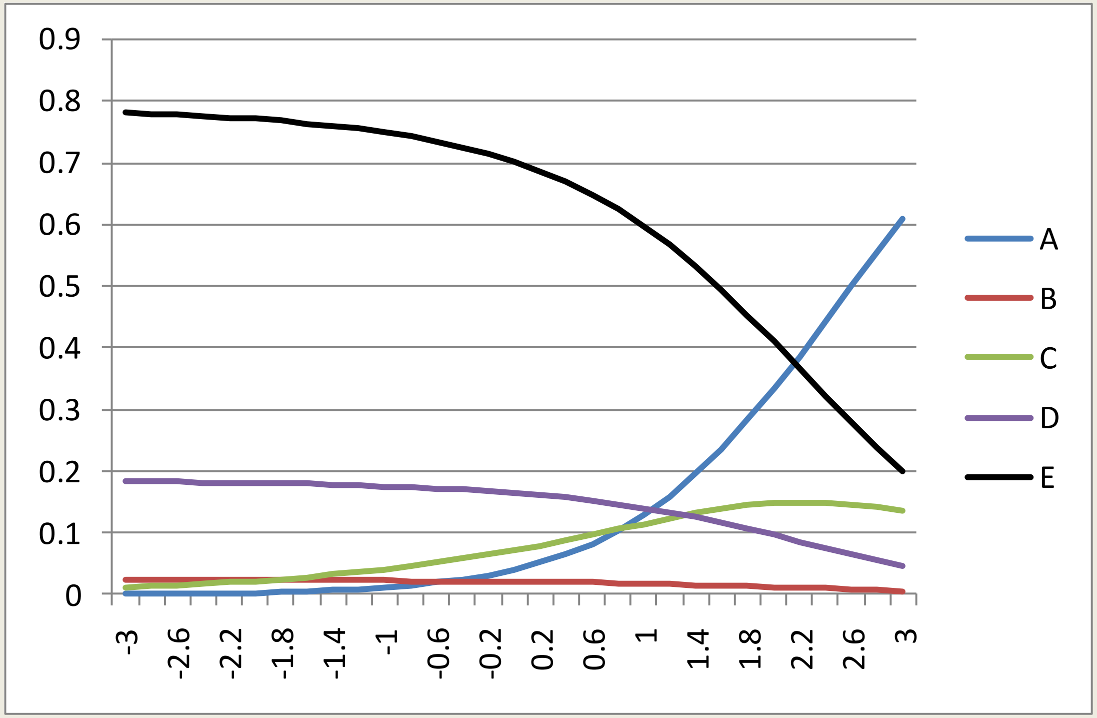
UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Nominal response curves

- ***"I prefer to eat"***

  (a) at home alone *(b) at home with others* *(c) in a restaurant*

  (d) at a friend's house *(e) doesn't matter*



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

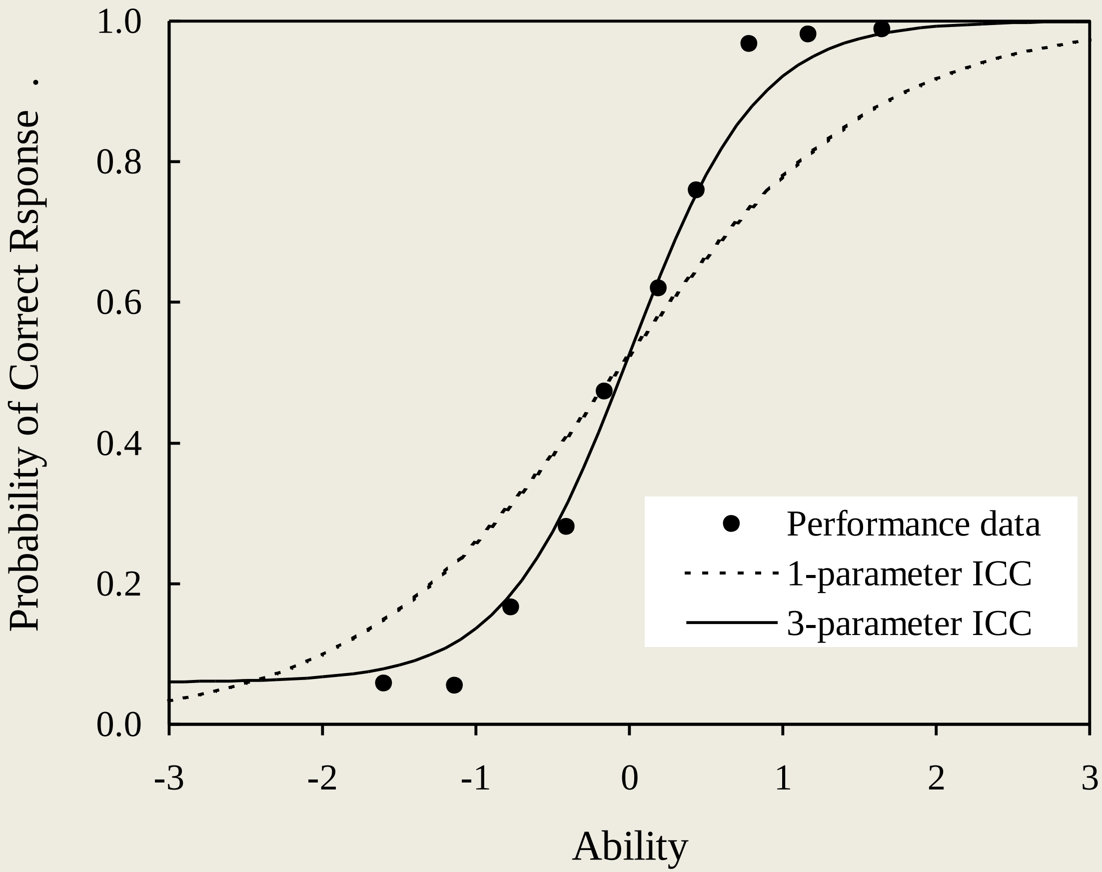
ASSESSING IRT MODEL FIT

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

IRT Model-Examinee Data Fit

- Assess model assumptions such as dimensionality
- Assess residuals and standardized residuals and examine consequences of model misfit (e.g., predicting score distributions)
- Check invariance properties (e.g., item bias)

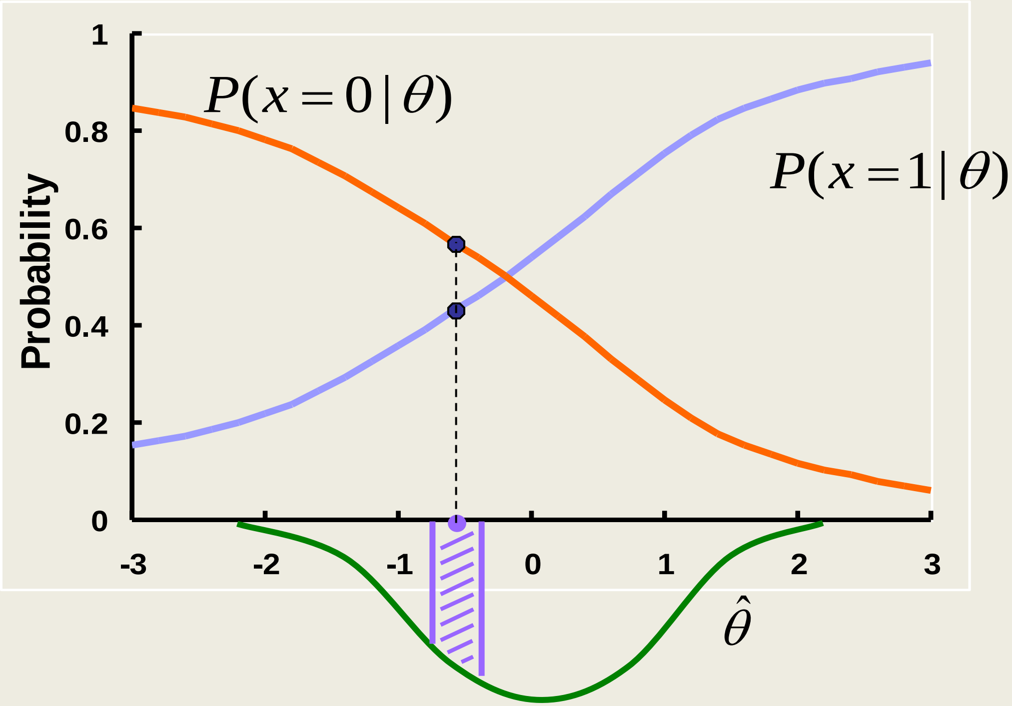
Does the model fit?



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Predicted vs. empirical binary data

- Divide the estimated distribution into ***k*** ability groups



$$P\left(x=0|\theta\right)+P\left(x=1|\theta\right)=1.0$$

IRT model fit

- R_{ij} is the raw residual of item i
  - where P-hat is the observed value, and P is expected

$$R_{ij} = \hat{P}_{ij} - P_{ij}$$

- SR_{ij} is the standardised residual

$$SR_{ij} = \frac{\hat{P}_{ij} - P_{ij}}{\sqrt{P_{ij}(1 - P_{ij}) / N_{ij}}}$$

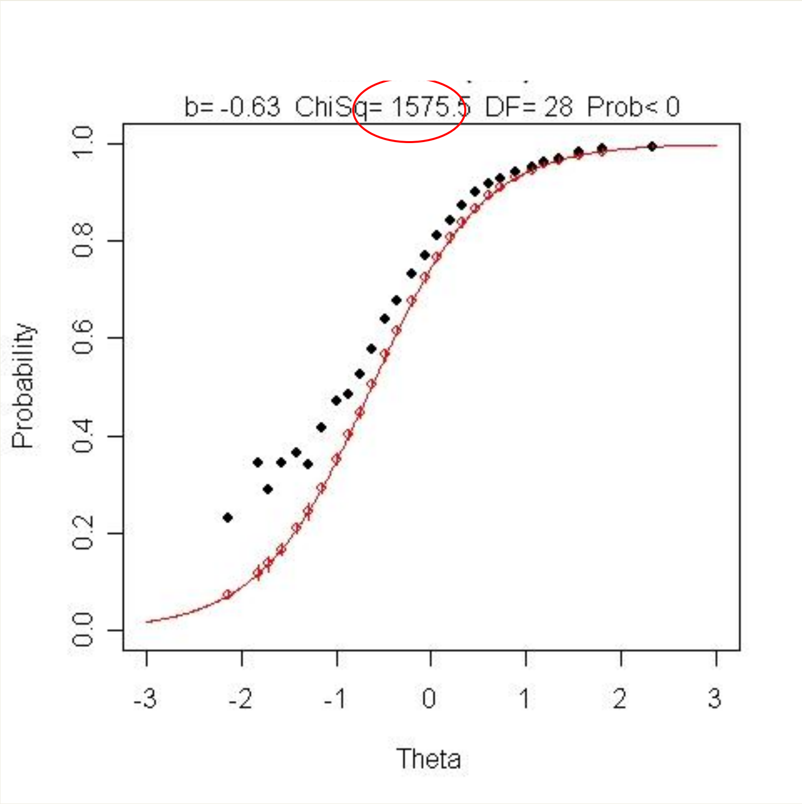
- k is the number of score categories

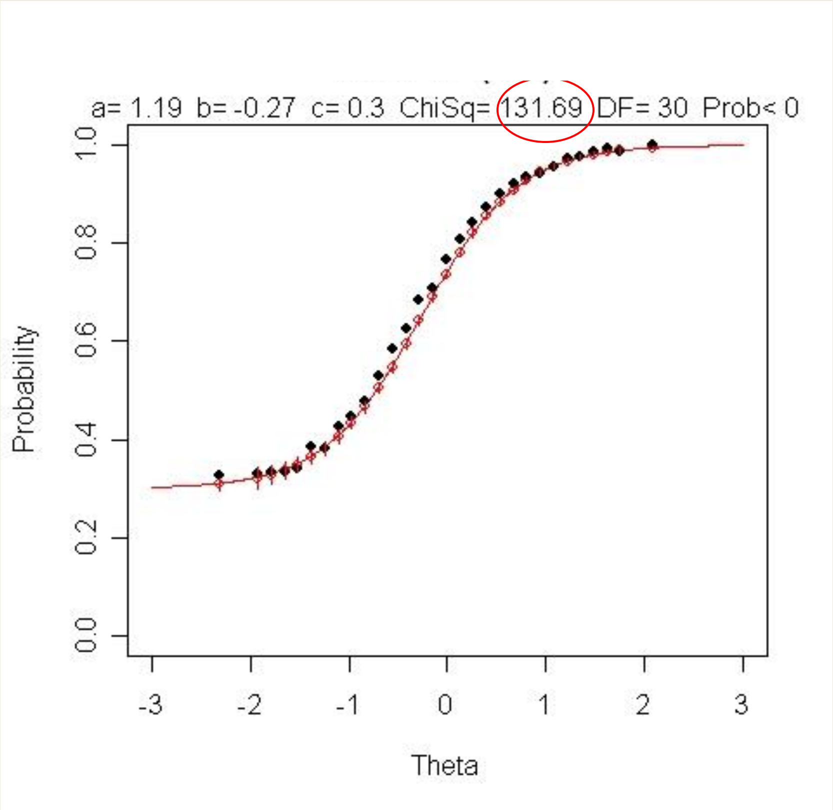
$$\chi_i^2 = \sum_{j=1}^{k} SR_{ij}^2$$

$$(df = k - \# \text{ item parameters in model})$$

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Fit Comparisons Under 3PL and 1PL Models





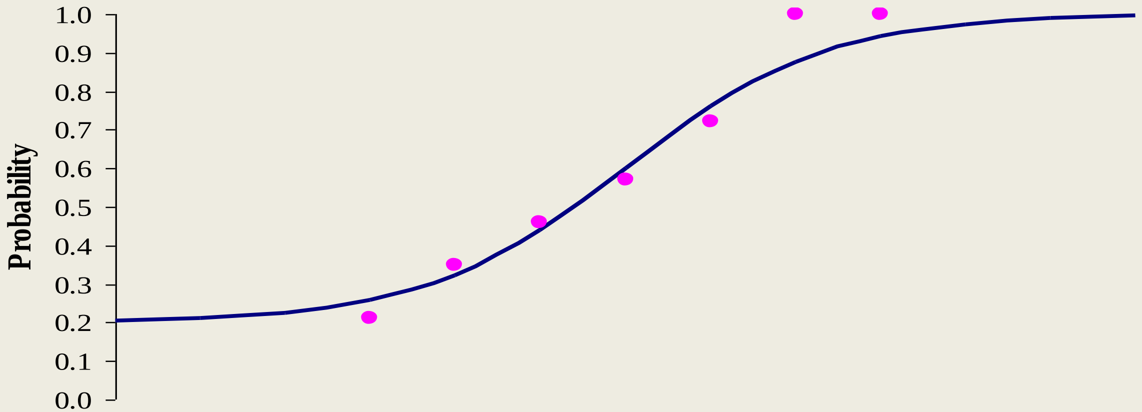
1PL **3PL**

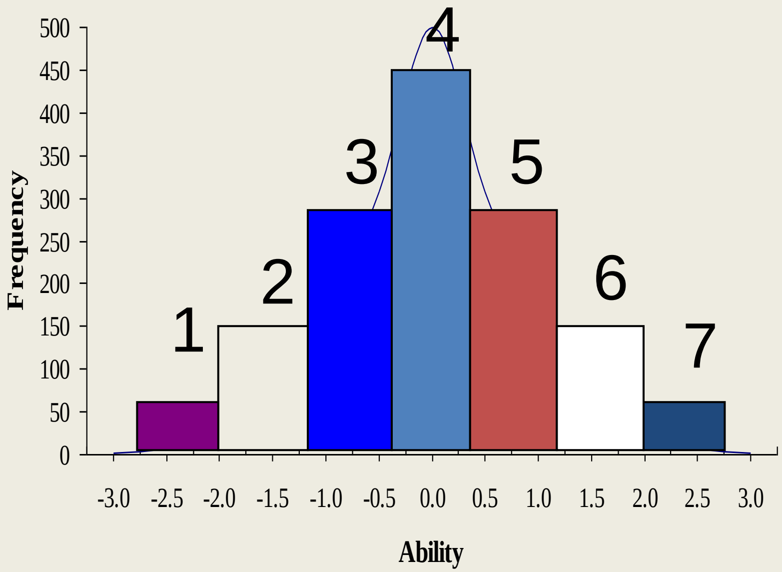
UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Calculating residuals

	Examinees												
Score group	1	2	3	4	5	6	7	8	9	10	P-hat	3PL	Res
1	1	0	0	0							0.25	0.287	-0.037
2	0	0	1	0	0	1					0.33	0.358	-0.028
3	1	0	1	0	1	0	1	0	0		0.44	0.465	-0.025
4	1	0	1	0	1	1	0	0	1	1	0.6	0.600	0.000
5	1	1	0	1	1	0	1	1	1		0.75	0.735	0.015
6	1	1	1	1	1	1					1	0.842	0.158
7	1	1	1	1							1	0.913	0.087

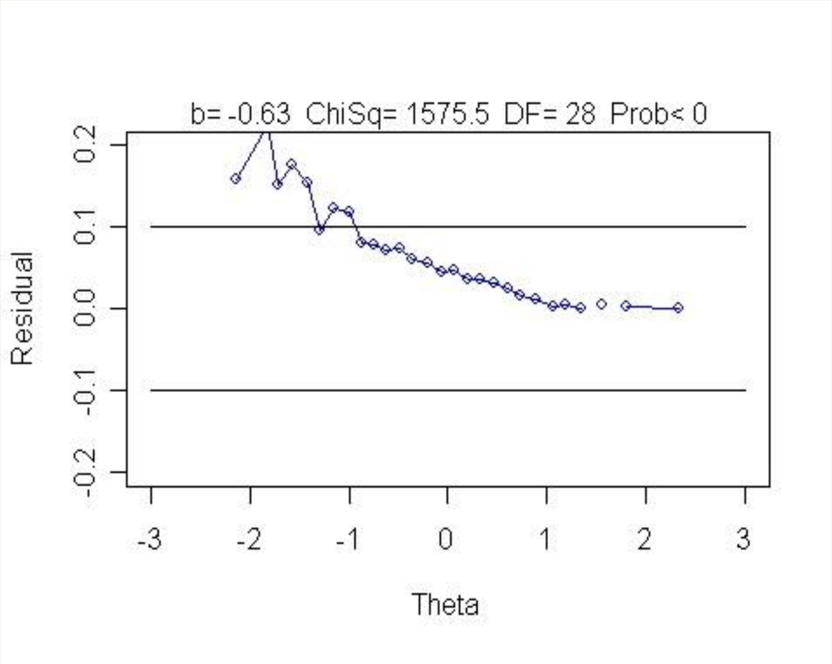
Plotting observed probabilities

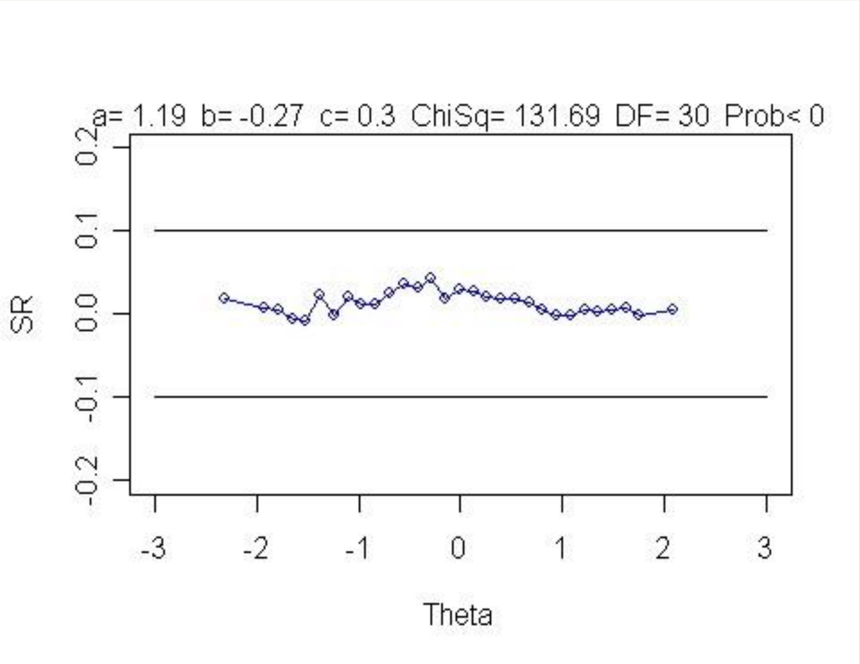




UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

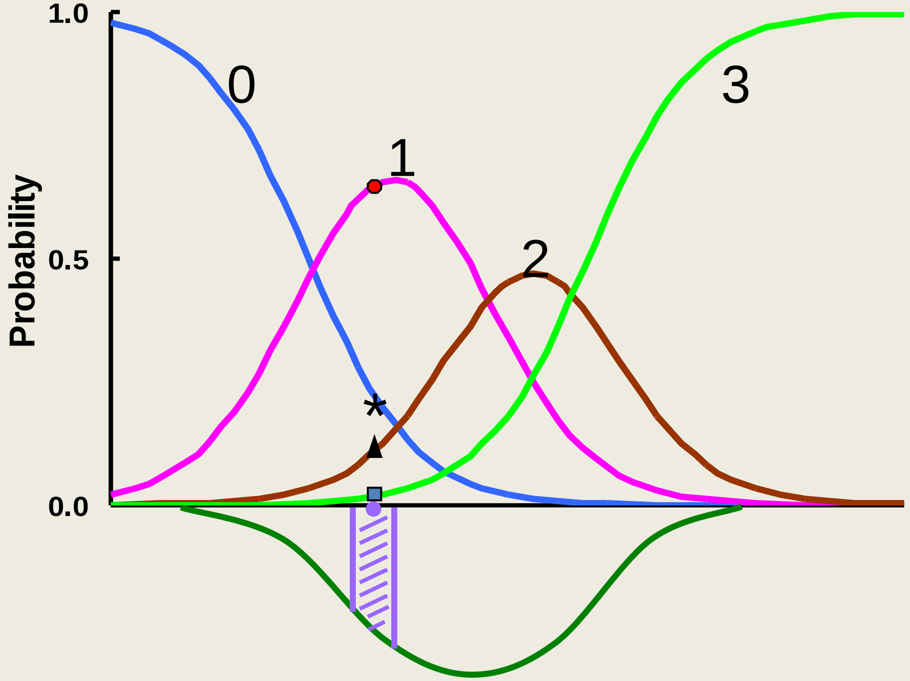
Fit Comparisons Under 3PL and IPL Models





1PL **3PL**

Predicted vs. empirical polytomous data



- For item ***i*** and score group ***j*** (***j=1…k***)

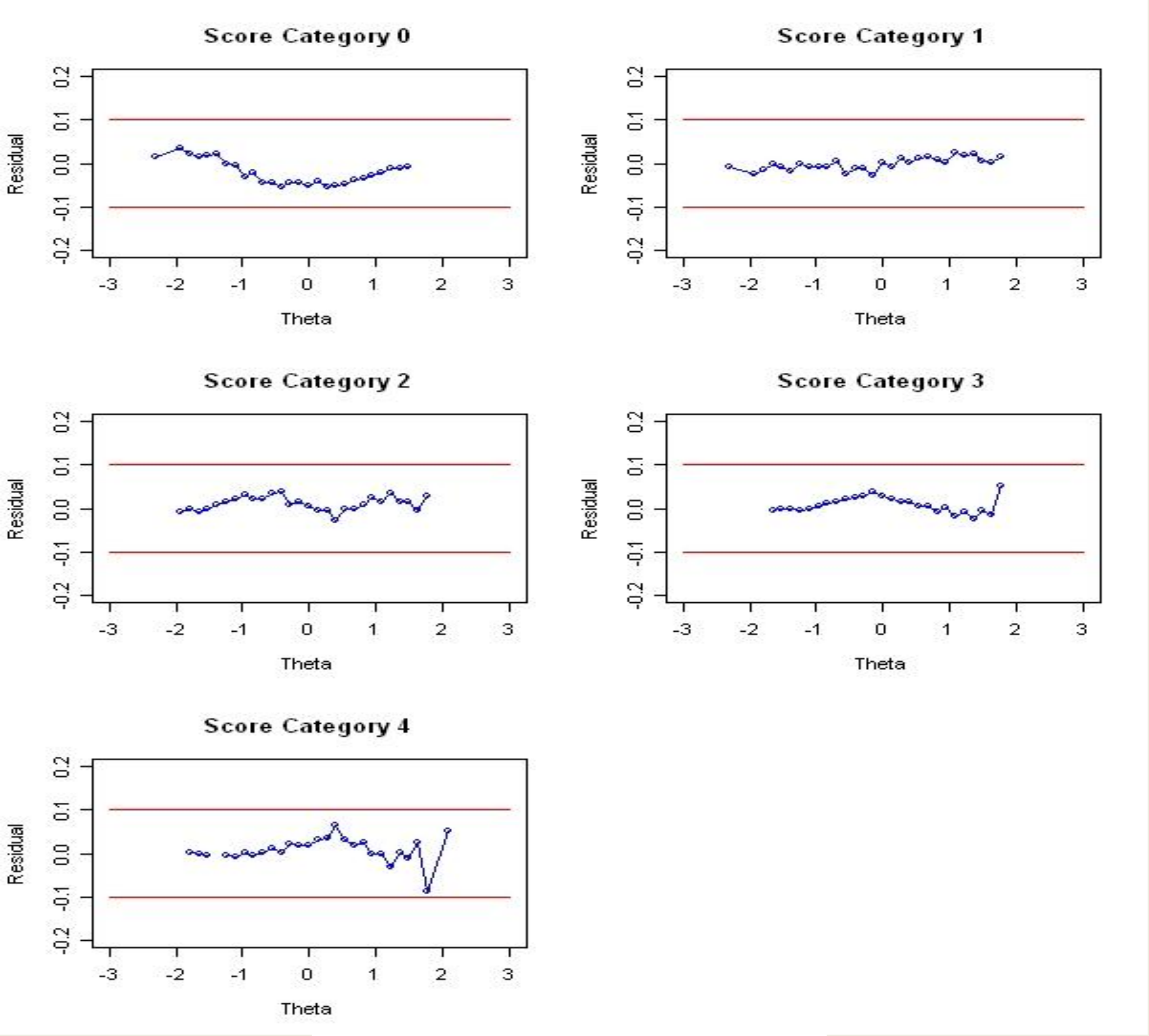
  ***N**$_{ij}$* = number of persons in j

  h is a response category

$$SR_{ijh} = \frac{\hat{P}_{ijh} - P_{ijh}}{\sqrt{P_{ijh}\left(1 - P_{ijh}\right)/N_{ij}}}$$

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Residual Plot for a Polytomous Item (GRM)



REVIEW OF IRT MODELS

How to choose from the many available IRT models?

- Is data binary, polytomous, or mixed?
- What is the psychological decision model/logic of responding?
- How large is sample size?
- How do model fit statistics compare?
  - Model fit results should be influential in model selection
- How much experience do I or my colleagues have with IRT models?
  - Or, can I get technical help?

Rasch vs. 2PL or 3PL Model?
(or PC vs. GR and GPCM?)

- This comparison has been of interest for many years, and generated quite emotional debate.
- Rasch model has many desirable properties
  - estimation of parameters is straightforward,
  - sample size does not need to be big,
  - number of items correct is the sufficient statistic for person's score,
  - measurement is completely additive,
  - specific objectivity (more on this tomorrow).
- But your data might not fit the Rasch model…

Rasch vs. 2PL or 3PL Model? (Cont.)

- Two-parameter logistic model is more complex
  - Often fits data better than the Rasch model
  - Requires larger samples (500+)
- Three-parameter logistic model is even more complex
  - Fits data where guessing is common better
  - Estimation is complex and estimates are not guaranteed without constraints
  - Sample needs to be large in applications.

Choice of model must be pragmatic

- Life is simple if the Rasch model suits your application and fits your data
- Desirable measurement properties of the Rasch model may make it a target model to achieve when constructing measures
  - Rasch maintained that if items have different discriminations, the latent trait is not unidimensional
- However, in many applications it is impossible to change the nature of the data
  - Take school exams with a lot of varied curriculum content to be squeezed in the test items
- There must be a pragmatic balance between the parsimony of the model and the complexity of the application

Coming in day 3…

- Rasch modelling!