CONCERTO

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

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

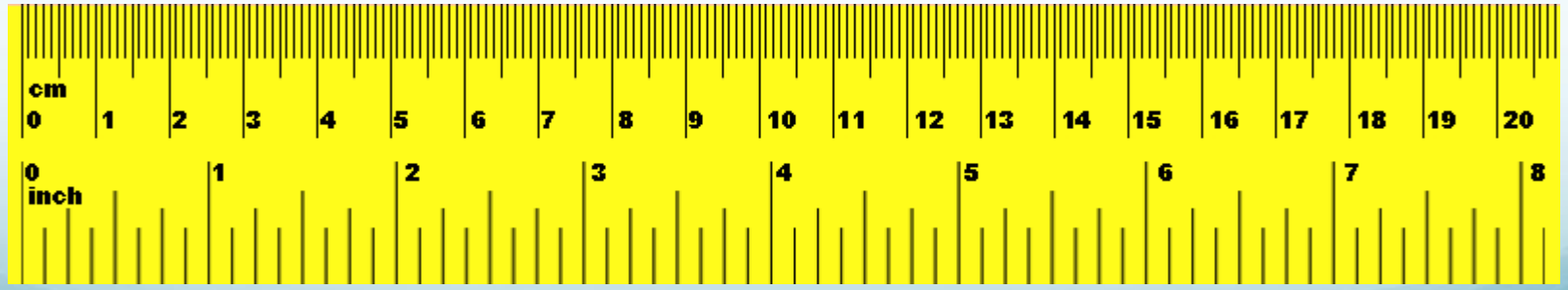
Learning outcomes

- 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

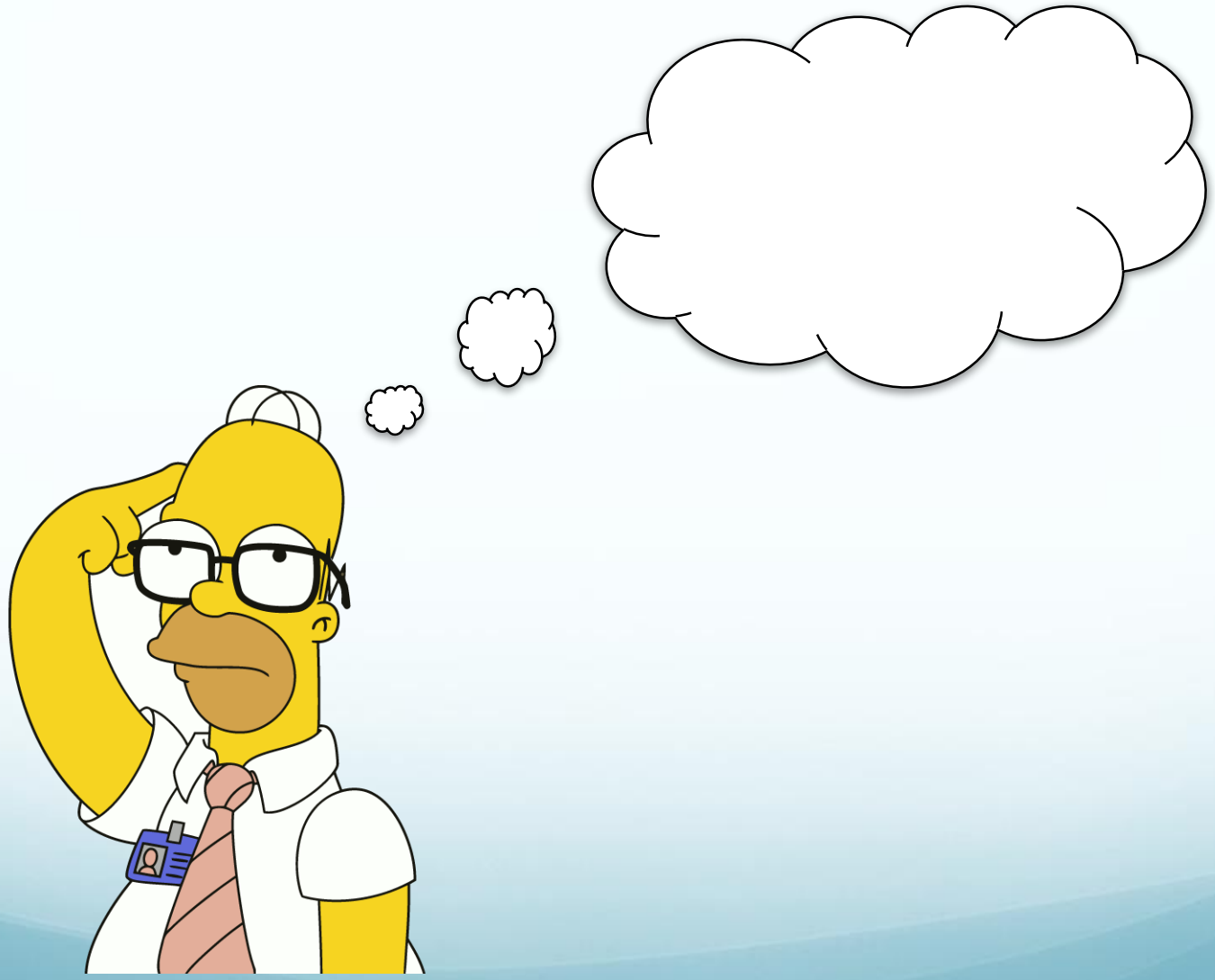
Measurement theory

- Psychometrics

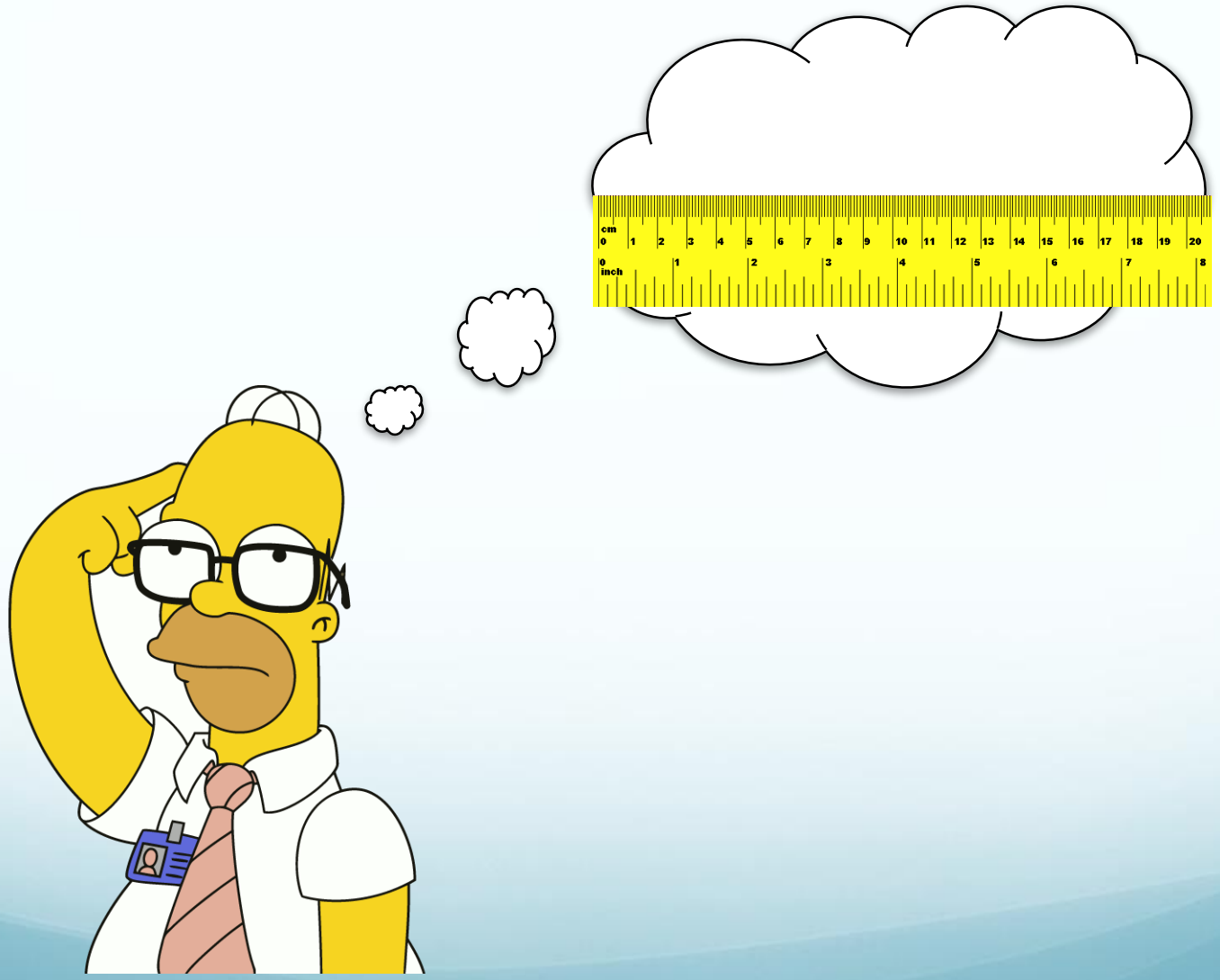
- Classical test theory
- Item response theory



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

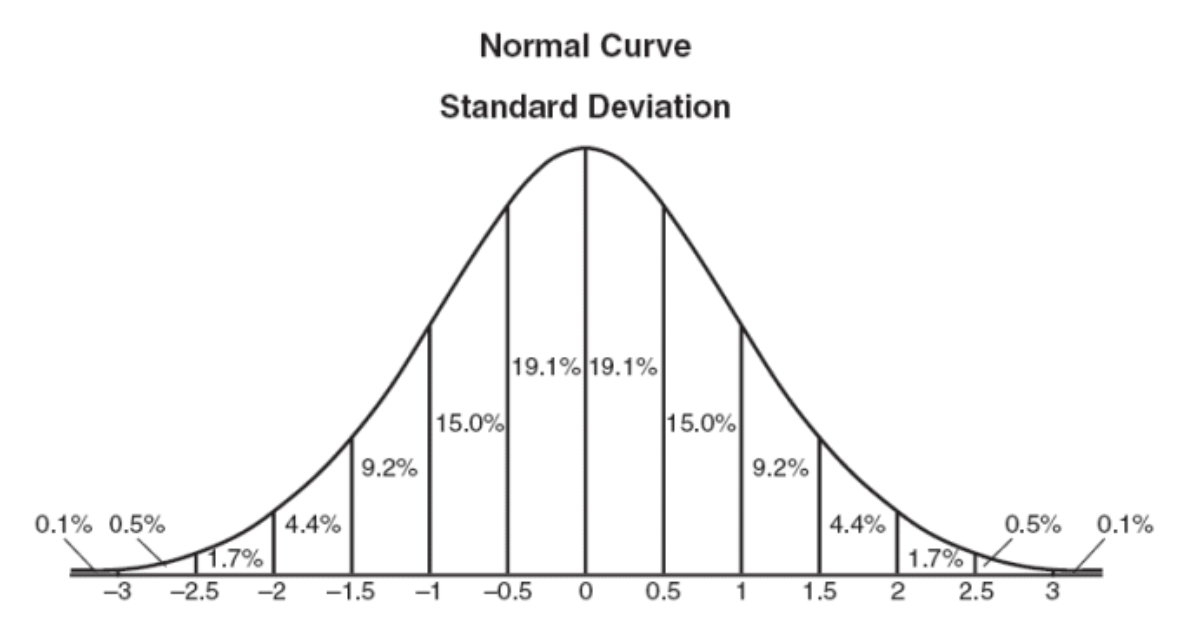
Validity

- 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)

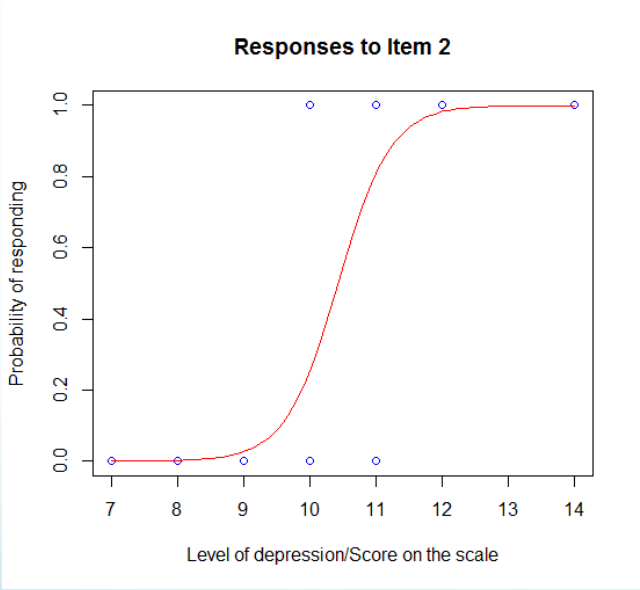


Calibration

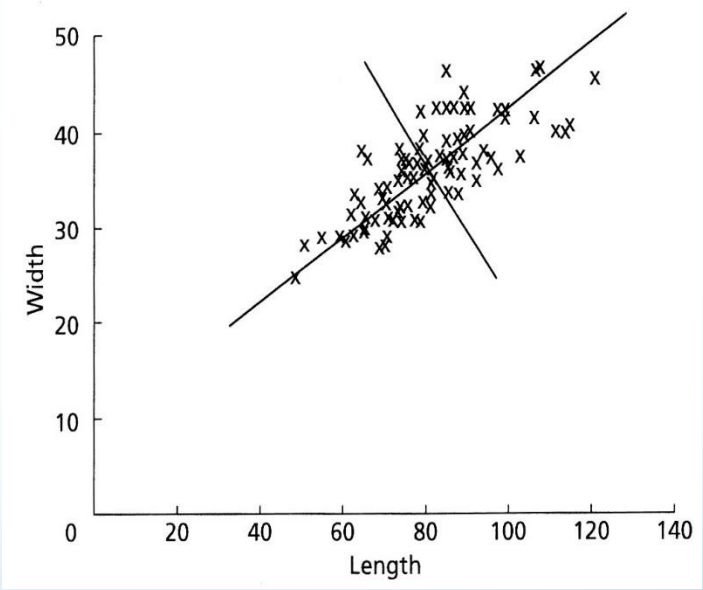
- Understanding scale scores in relation to other constructs, educational levels, symptoms.
- Receiver operating characteristics (ROC curves)
- Co-calibration with other scales using IRT

Psychometric approaches

Item response theory

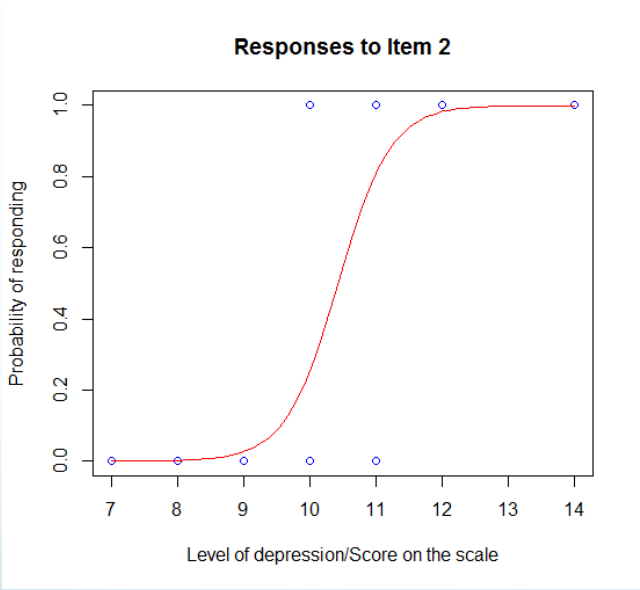


Classical test theory

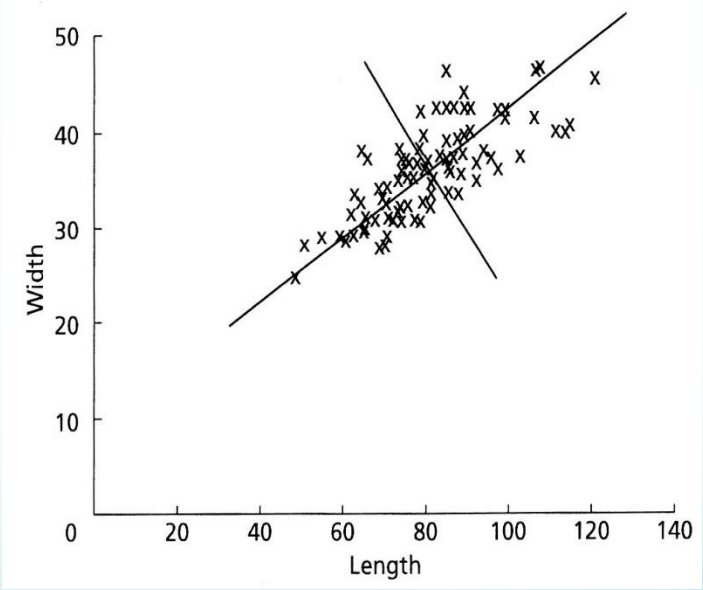


Psychometric approaches

Probabilistic

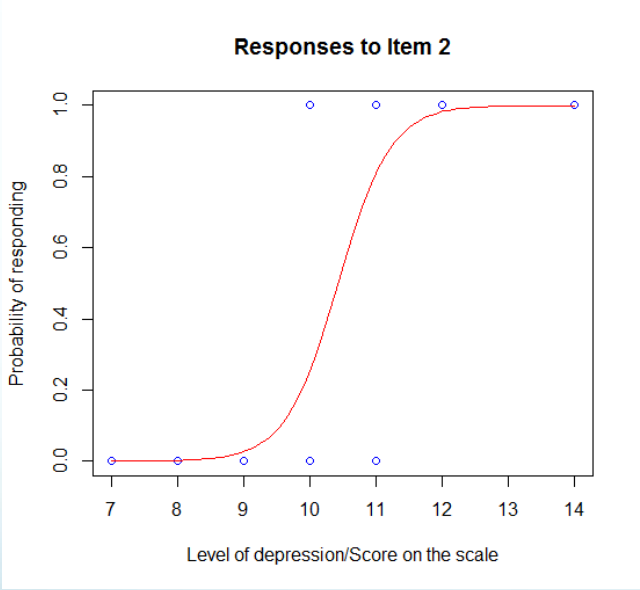


Correlational

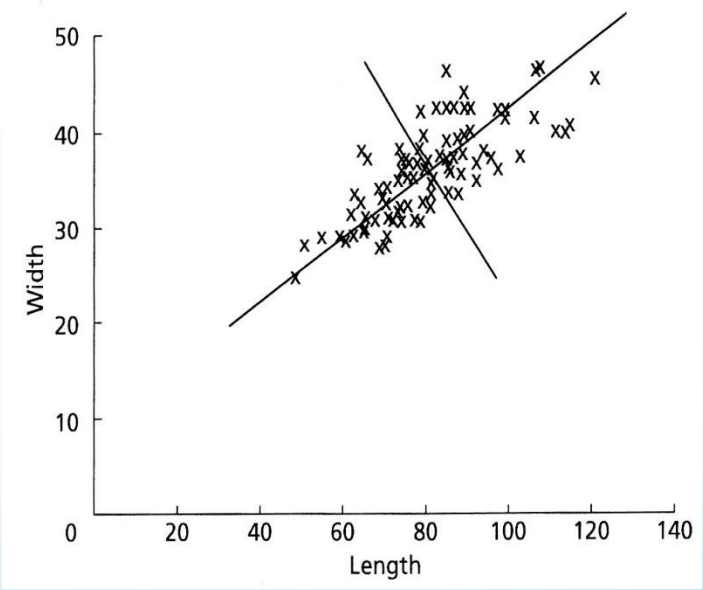


Psychometric approaches

Modern



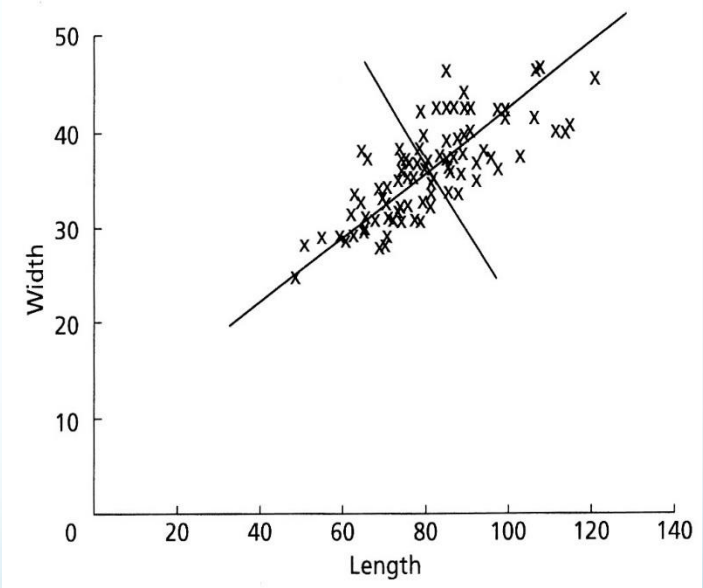
'Classic'



Psychometric approaches

'Hard'

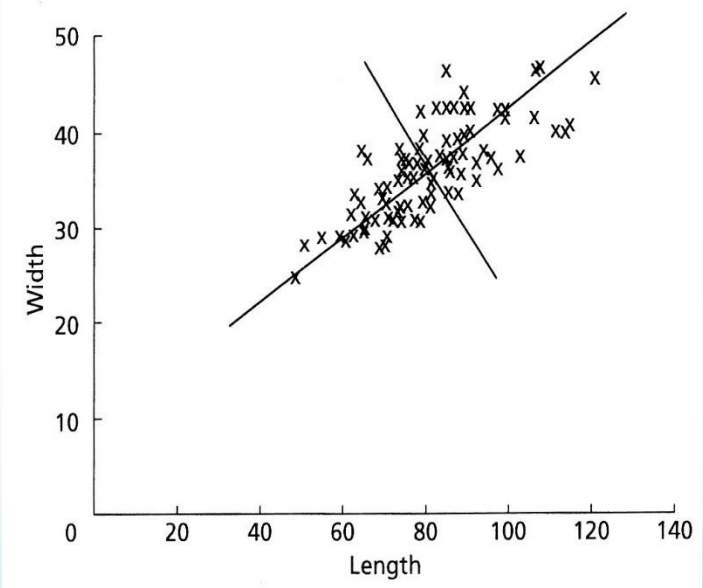
'Easy'



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:

1	Some days I feel unhappy	Agree	Disagree
2	I don't enjoy things anymore	Agree	Disagree
3	I don't laugh anymore	Agree	Disagree
4	I want to die	Agree	Disagree
5	Sometimes I'm sad	Agree	Disagree
6	Life is difficult right now	Agree	Disagree
7	Things won't get better	Agree	Disagree

Probabilistic approach to responding to items

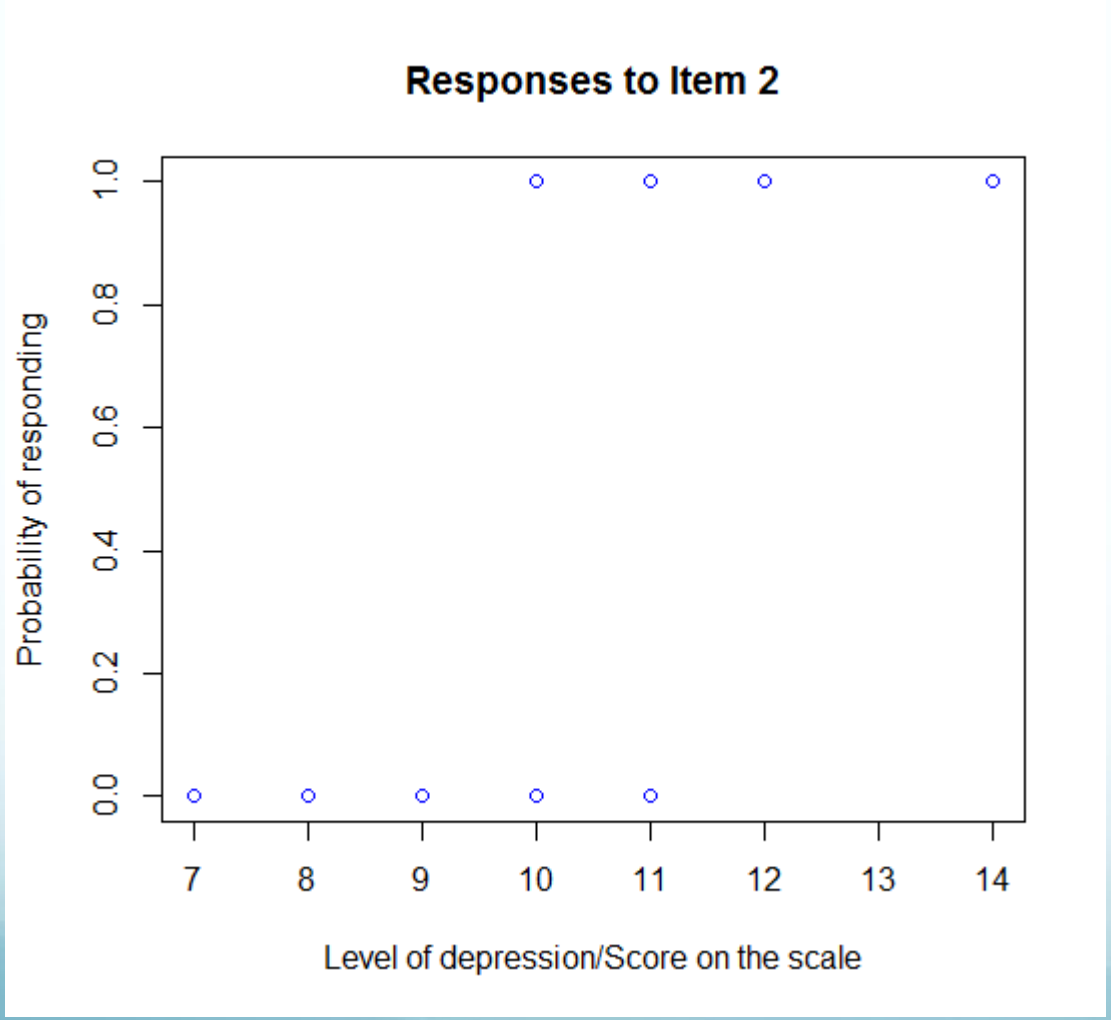
Consider the following depression questionnaire:

1	Some days I feel unhappy	Agree	Disagree
2	I don't enjoy things anymore	Agree	Disagree
3	I don't laugh anymore	Agree	Disagree
4	I want to die	Agree	Disagree
5	Sometimes I'm sad	Agree	Disagree
6	Life is difficult right now	Agree	Disagree
7	Things won't get better	Agree	Disagree

We can score Agree = 1 and Disagree = 0

Probabilistic approach to responding to items

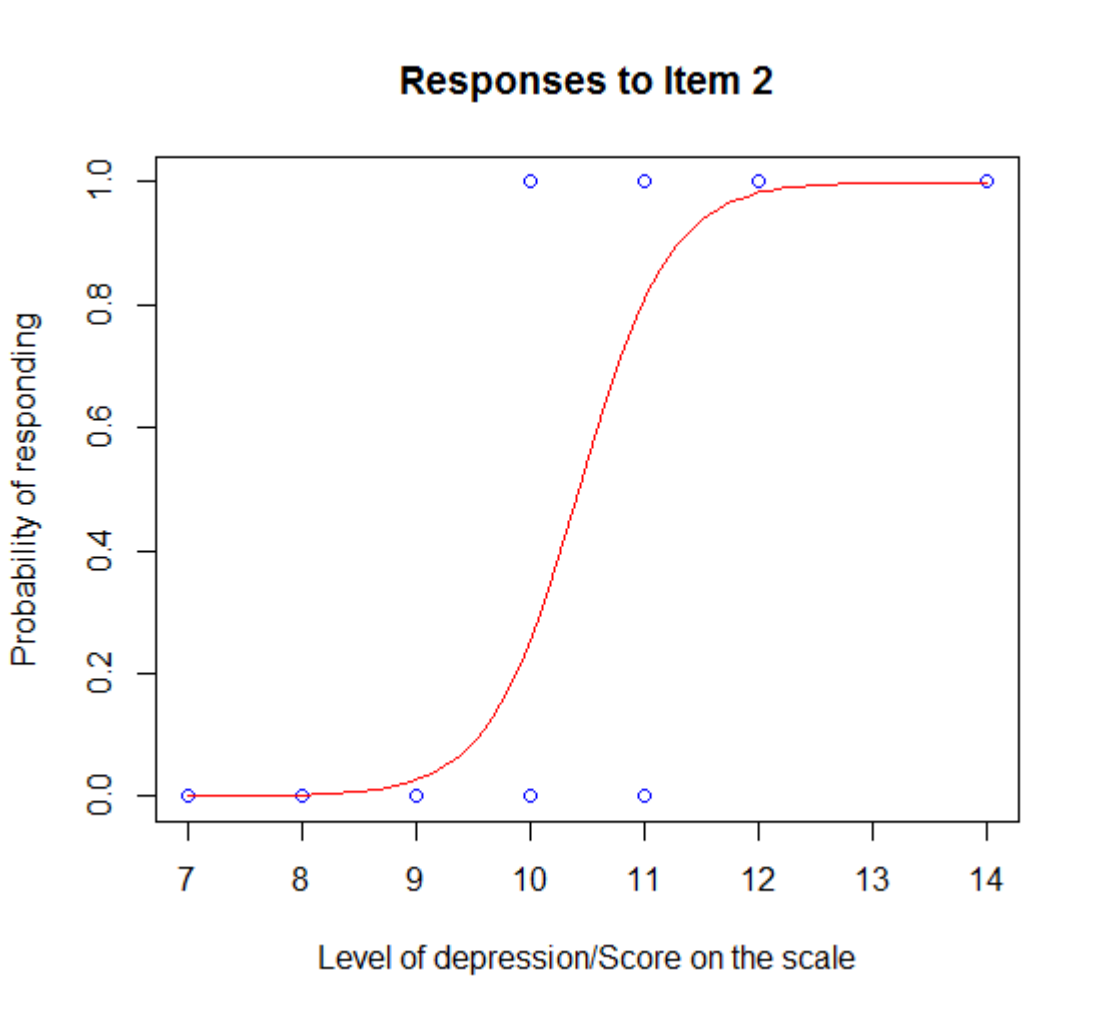
2	I don't enjoy things anymore	Agree	Disagree



Person	Level of depression	Agree?
1	6	0
2	7	0
3	7	0
4	8	0
5	9	0
6	9	0
7	9	0
8	9	0
9	9	1
10	10	0
11	10	0
12	10	1
13	10	0
14	10	1
15	11	1
16	11	1
17	11	1
18	12	1
19	12	1
20	14	1

Probabilistic approach to responding to items

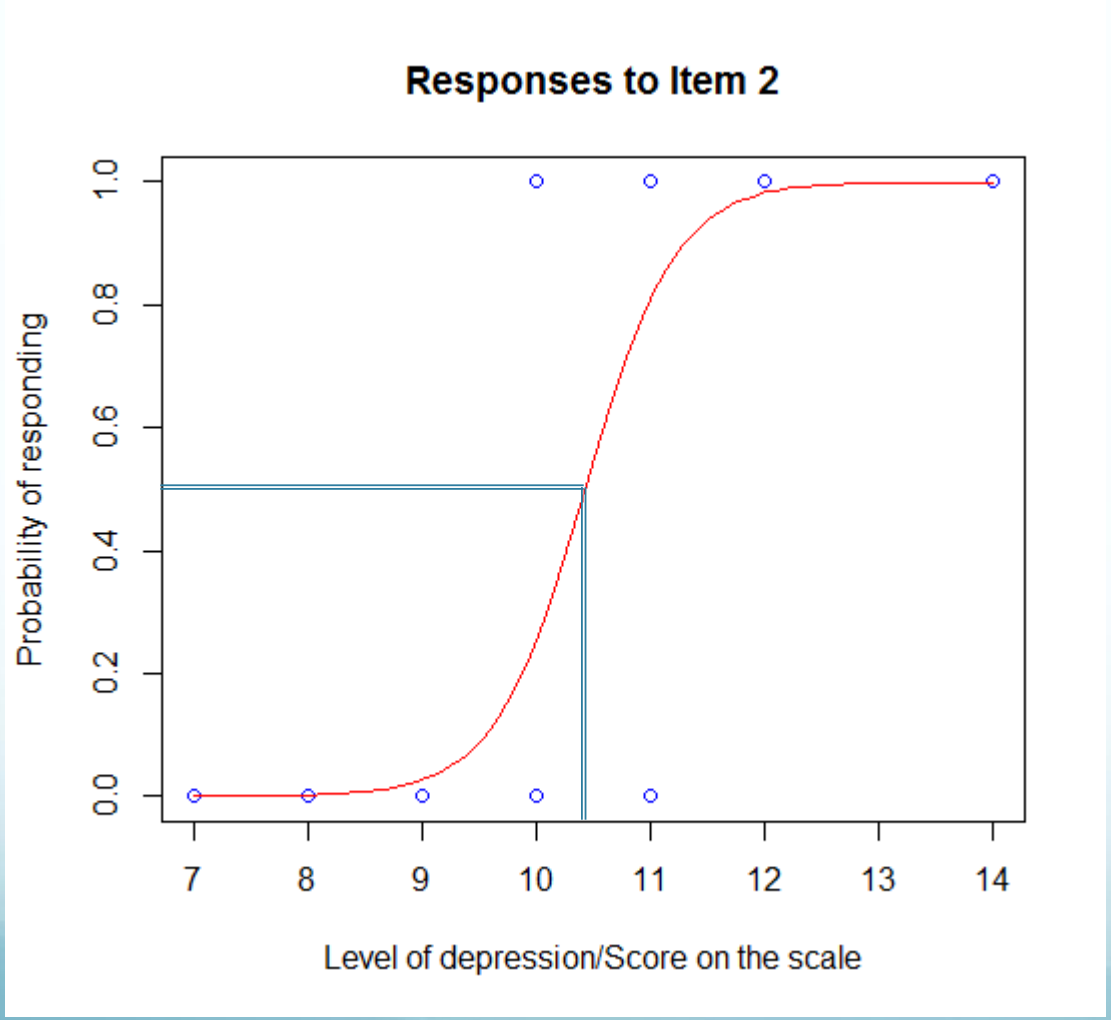
2	I don't enjoy things anymore	Agree	Disagree



Person	Level of depression	Agree?
1	6	0
2	7	0
3	7	0
4	8	0
5	9	0
6	9	0
7	9	0
8	9	0
9	9	1
10	10	0
11	10	0
12	10	1
13	10	0
14	10	1
15	11	1
16	11	1
17	11	1
18	12	1
19	12	1
20	14	1

Plotting item difficulty and person ability

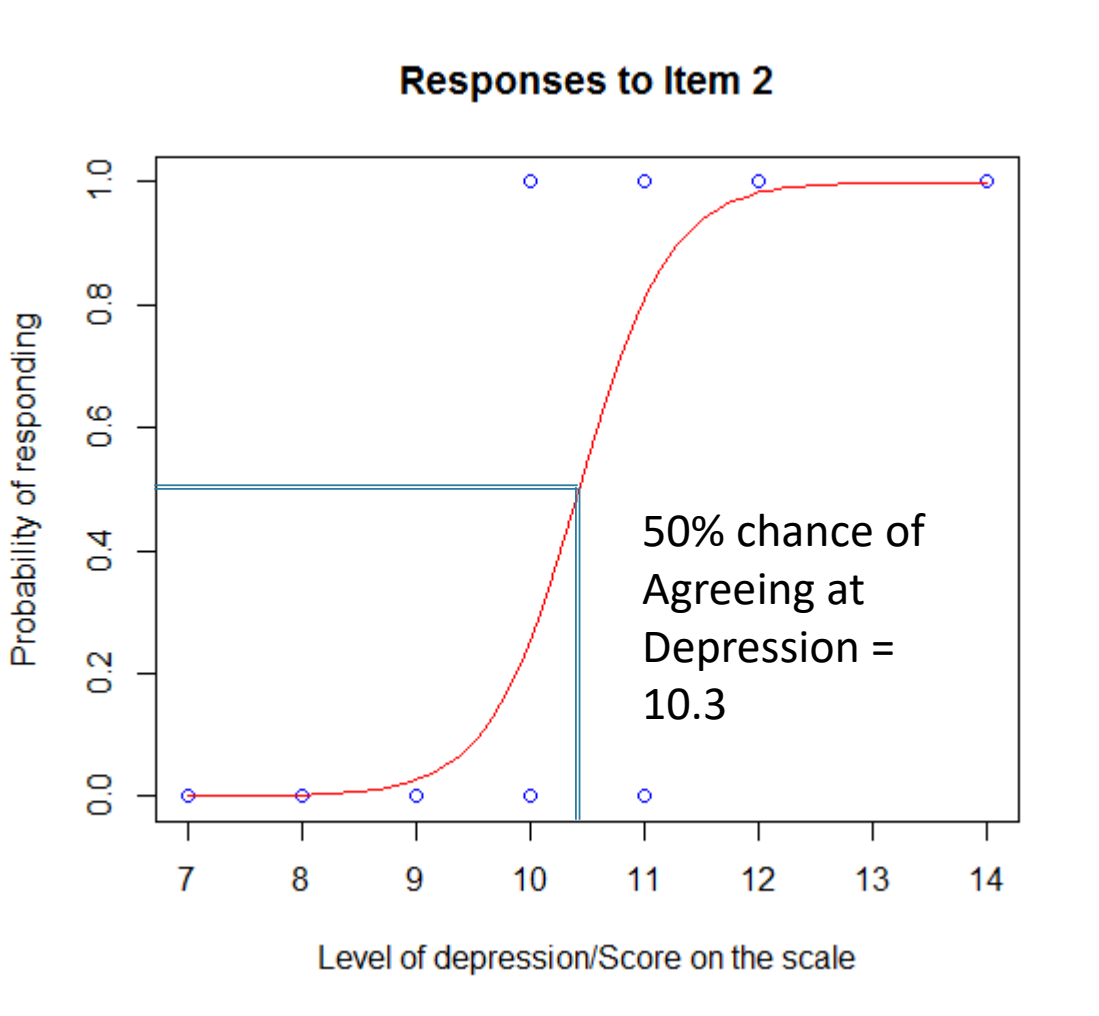
2	I don't enjoy things anymore	Agree	Disagree



Person	Level of depression	Agree?
1	6	0
2	7	0
3	7	0
4	8	0
5	9	0
6	9	0
7	9	0
8	9	0
9	9	1
10	10	0
11	10	0
12	10	1
13	10	0
14	10	1
15	11	1
16	11	1
17	11	1
18	12	1
19	12	1
20	14	1

Plotting item difficulty and person ability

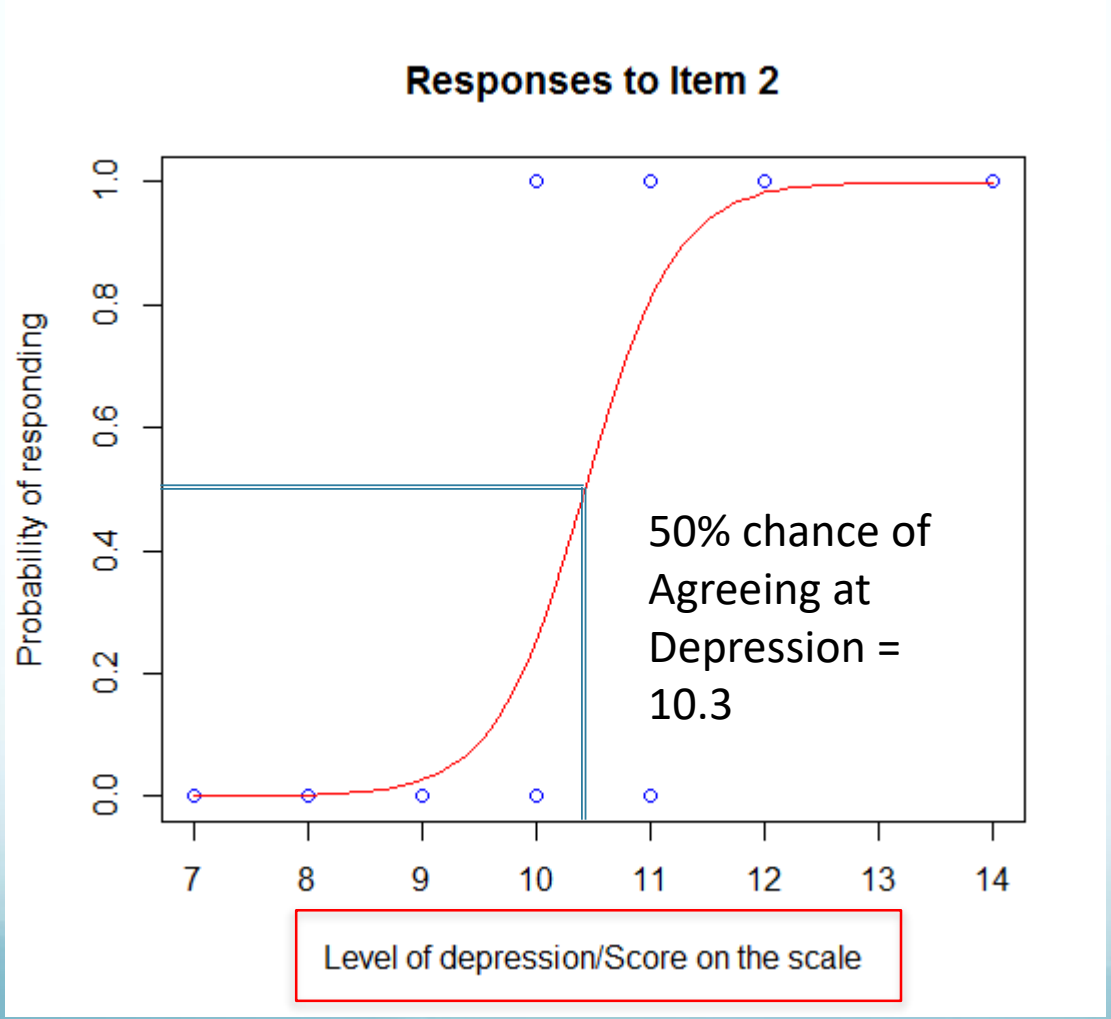
2	I don't enjoy things anymore	Agree	Disagree



Person	Level of depression	Agree?
1	6	0
2	7	0
3	7	0
4	8	0
5	9	0
6	9	0
7	9	0
8	9	0
9	9	1
10	10	0
11	10	0
12	10	1
13	10	0
14	10	1
15	11	1
16	11	1
17	11	1
18	12	1
19	12	1
20	14	1

Plotting item difficulty and person ability

2	I don't enjoy things anymore	Agree	Disagree



Person	Level of depression	Agree?
1	6	0
2	7	0
3	7	0
4	8	0
5	9	0
6	9	0
7	9	0
8	9	0
9	9	1
10	10	0
11	10	0
12	10	1
13	10	0
14	10	1
15	11	1
16	11	1
17	11	1
18	12	1
19	12	1
20	14	1

Level of depression

Fatigue

Neuroticism

IQ

Anxiety

Quality of life

Health Behaviour

Religeousness

Theta (θ)

Theta (θ)

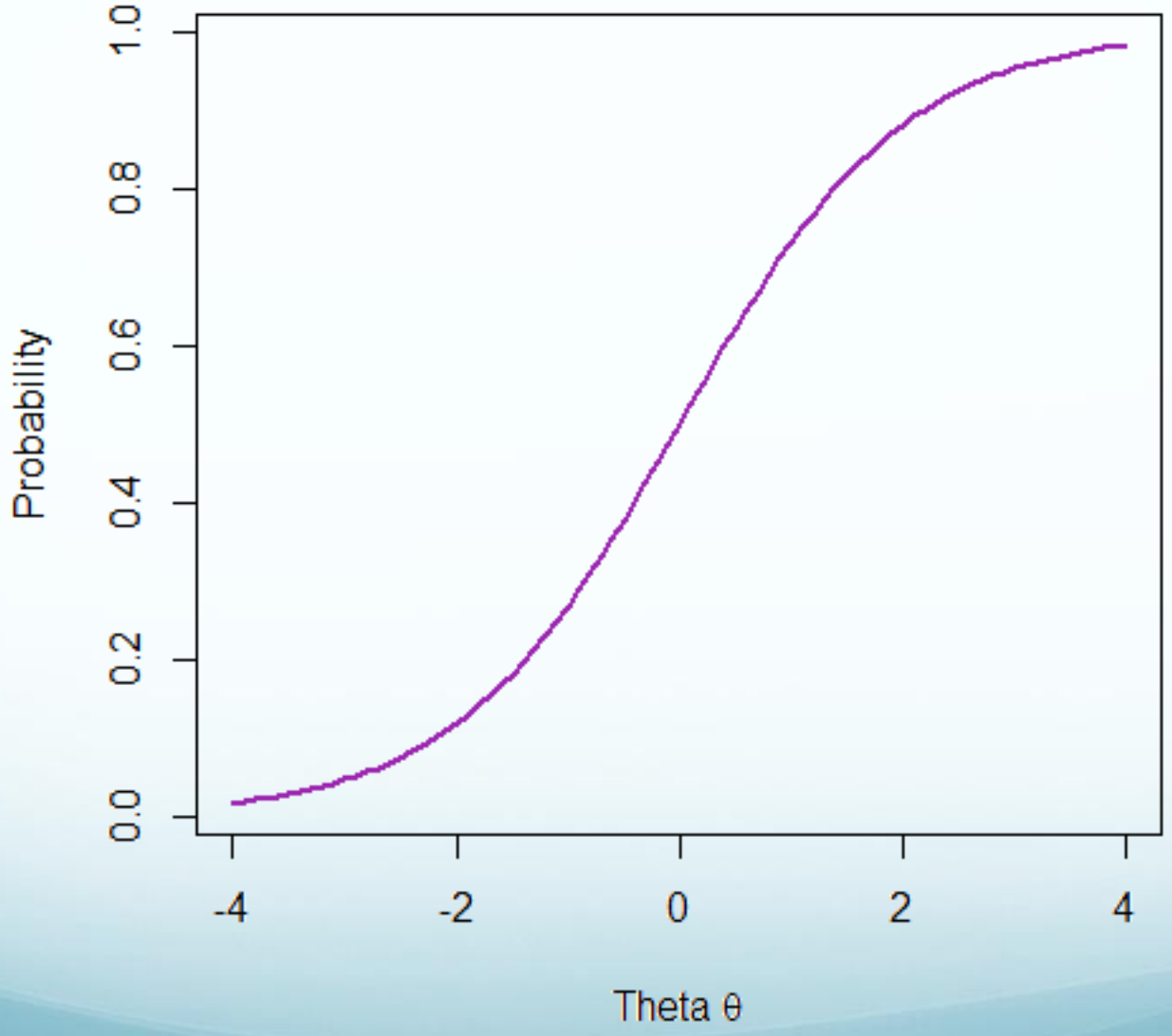
A common metric on which to talk about

- 'Difficulty' of items
- 'Ability' of persons

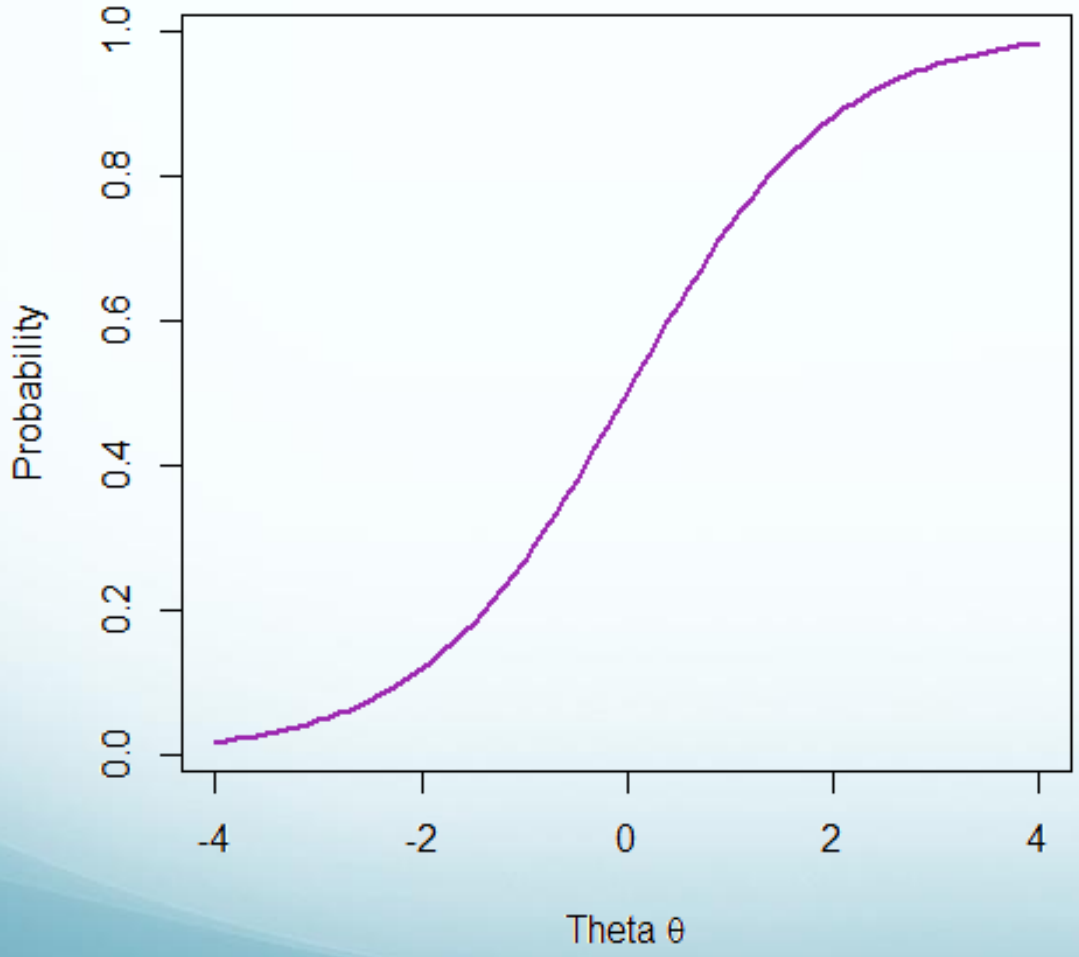
Classical approach versus Item Response Theory

	Classical	IRT
Modelling / Interpretation	Total score	Individual items (questions)
Accuracy / Information	Same for all participants and scores	Estimated for each score / participant
Adaptivity	Virtually not possible	Possible
Score	Depends on the items	Item independent
Item Parameters	Sample dependent	Sample independent
Preferred items	Average difficulty	Any difficulty

Item response function



A maths question



What is 4 x 8?

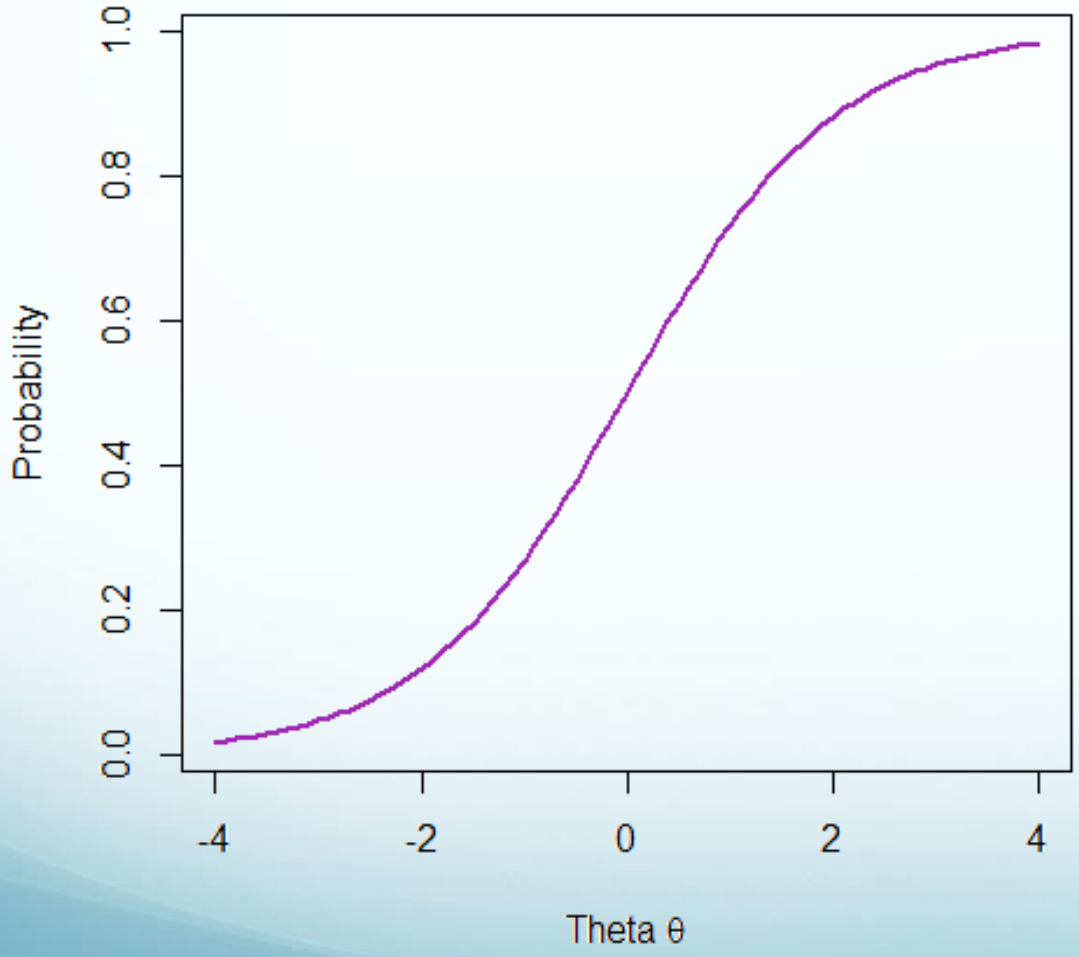
a.) 20

b.) 32

c.) 40

d.) 16

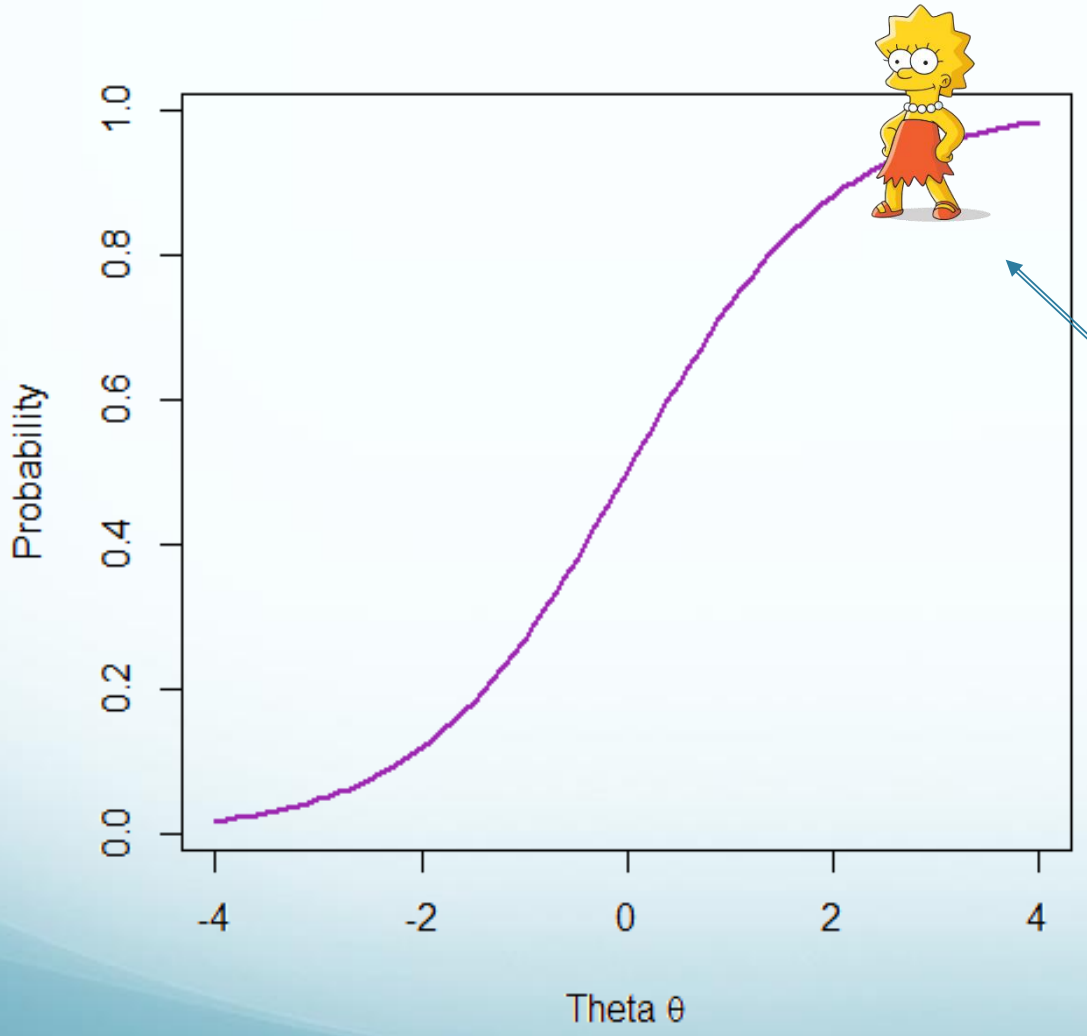
A maths question



What is 4 x 8?

Candidate One

A maths question

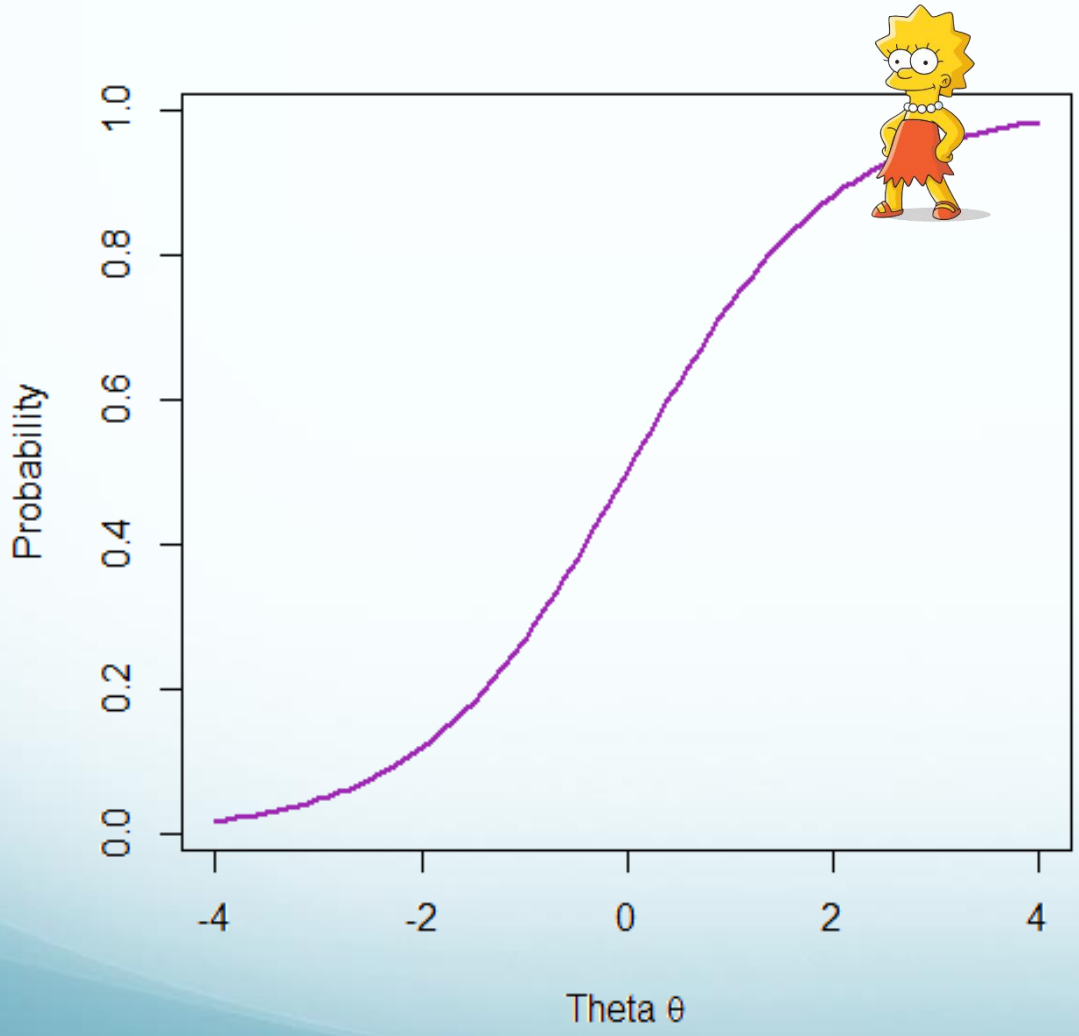


What is 4 x 8?

Candidate One

Probability close to 1
Person ability = 2.5

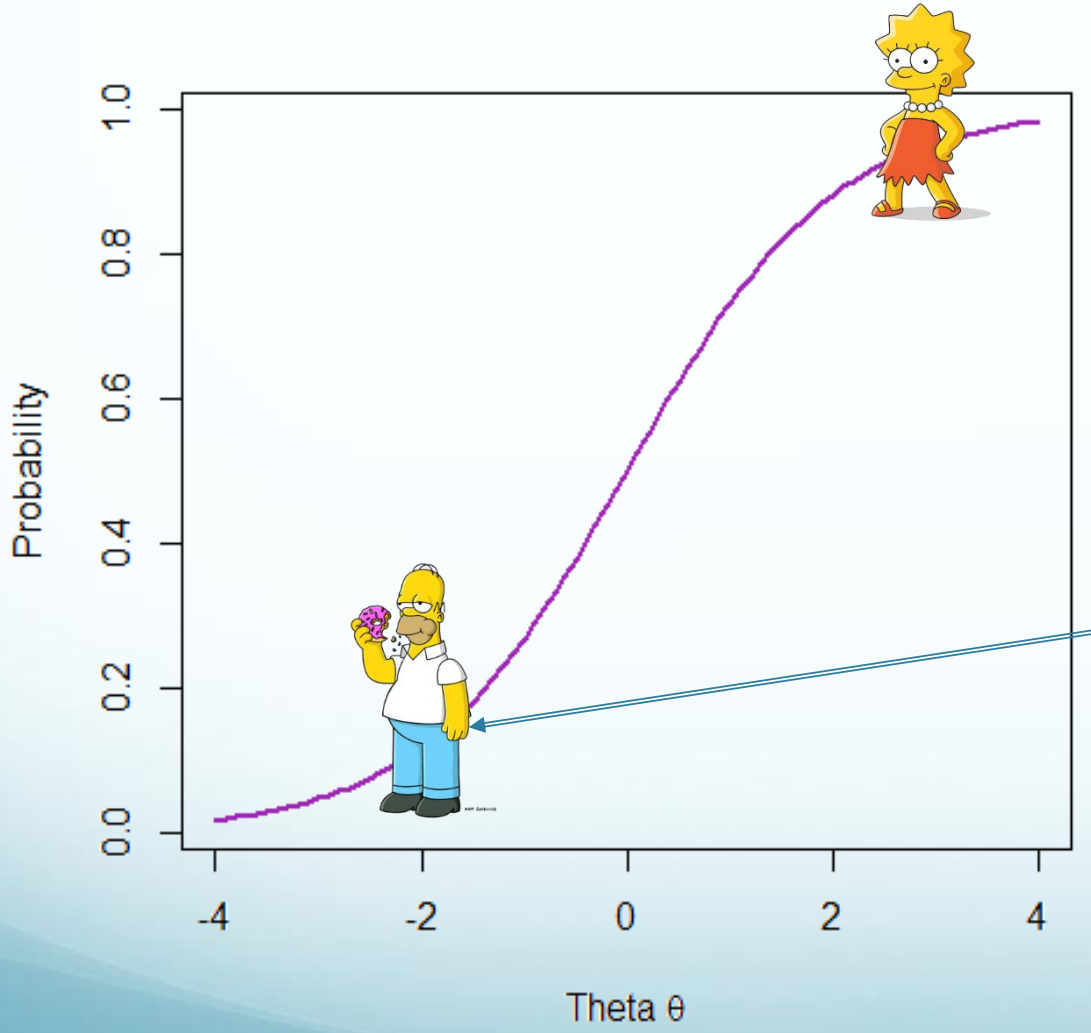
A maths question



What is 4 x 8?

Candidate Two

A maths question

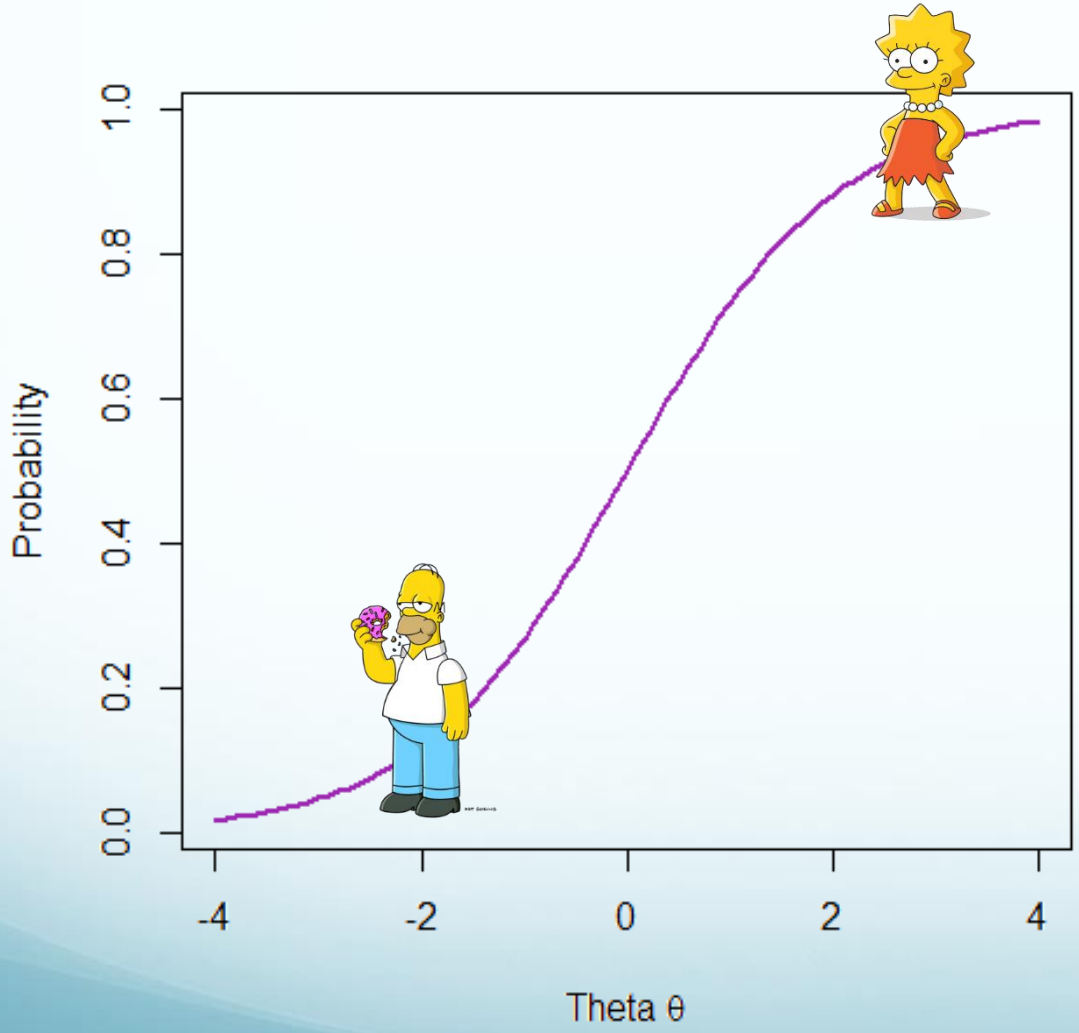


What is 4 x 8?

Candidate Two

Probability close to .1
Person ability = -2

A maths question

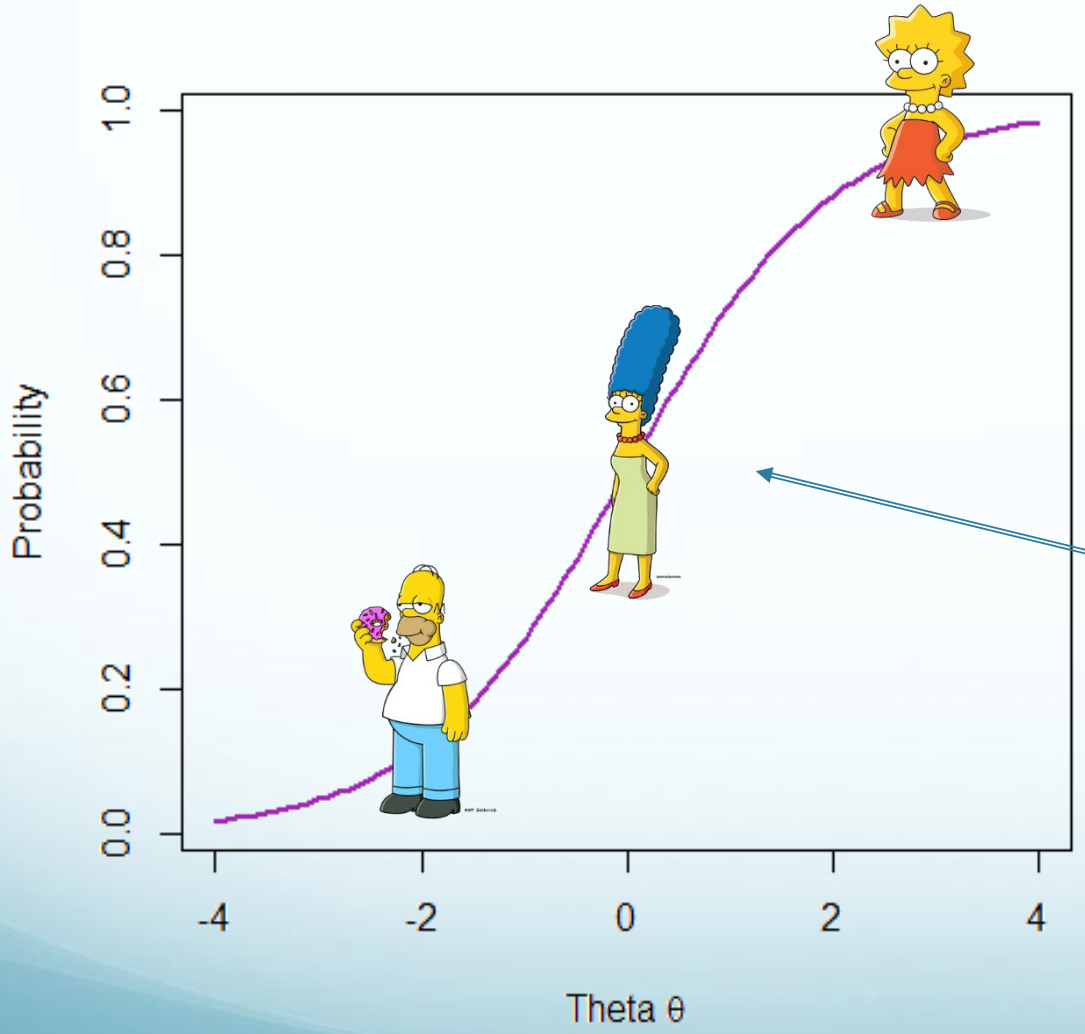


What is 4 x 8?

Candidate Three

Probability close to .5
Person ability = 0

A maths question



What is 4 x 8?

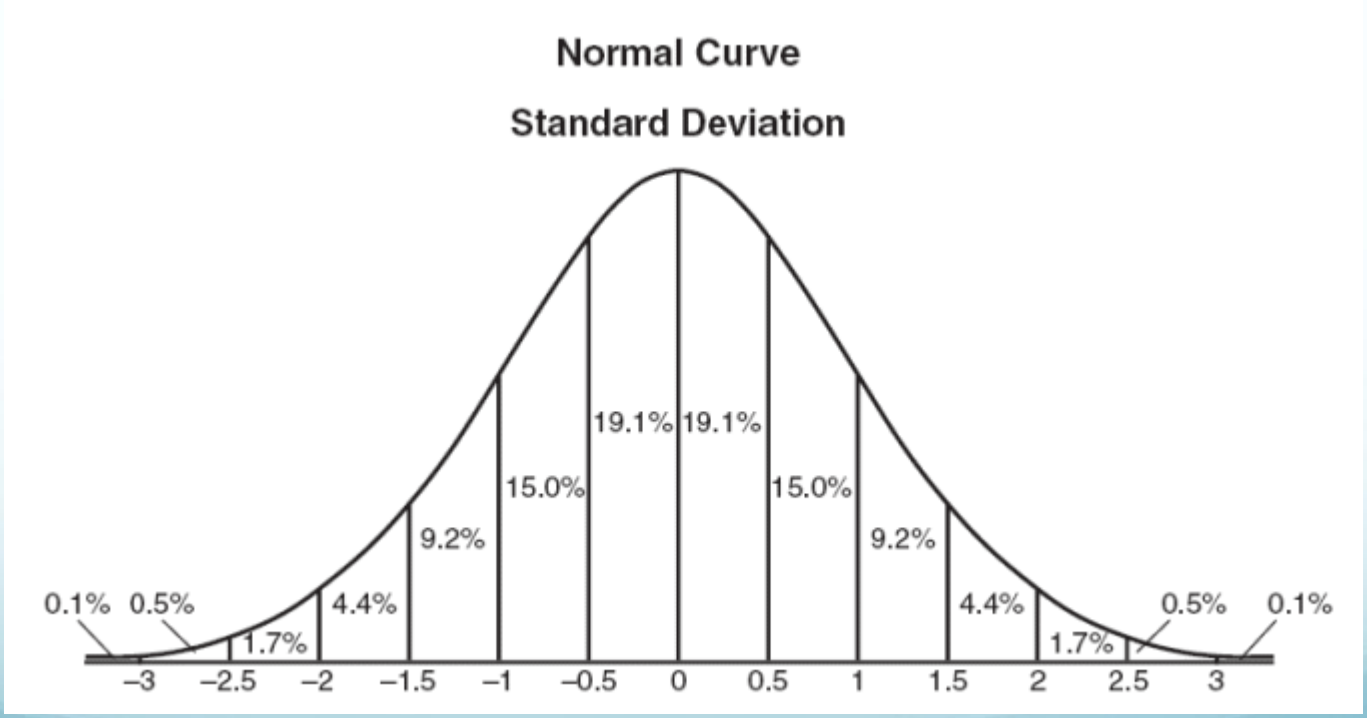
Candidate Three

Probability close to .5
Person ability = 0

A note about theta

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

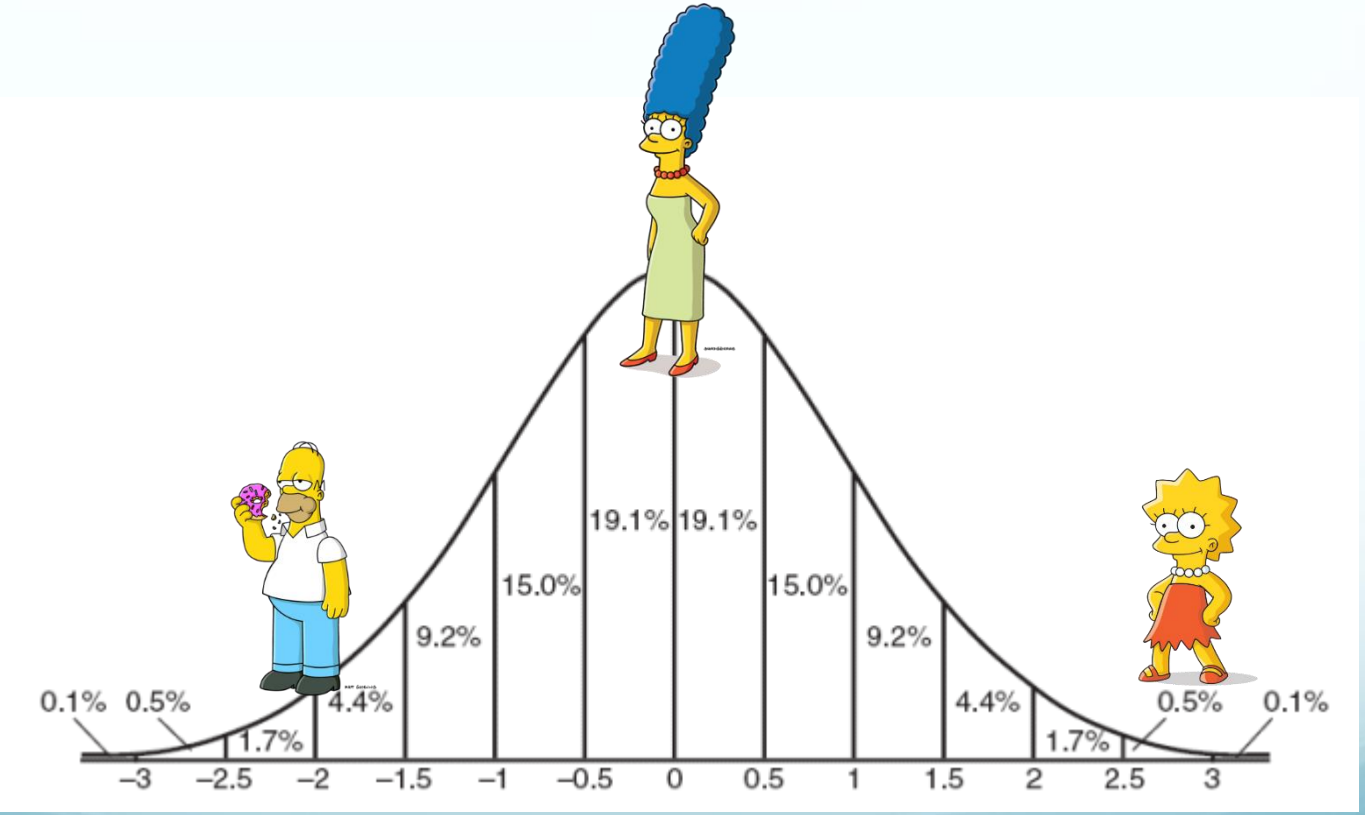
Where theta = 0 is the population mean then ± 1 theta = ± 1 standard deviation



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 ± 1 theta = ± 1 standard deviation



Types of IRT model

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

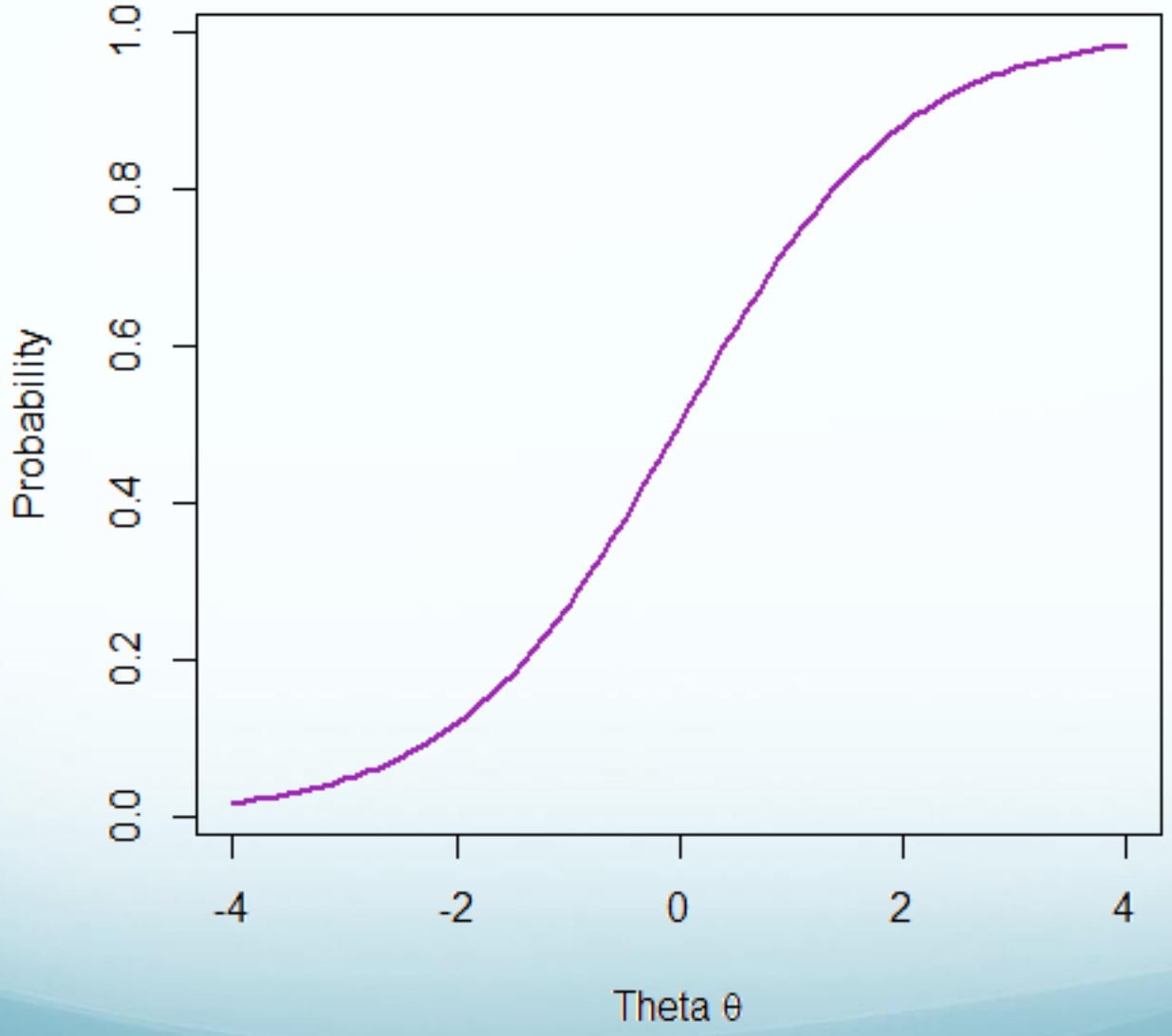
Question

What does theta (θ) represent?

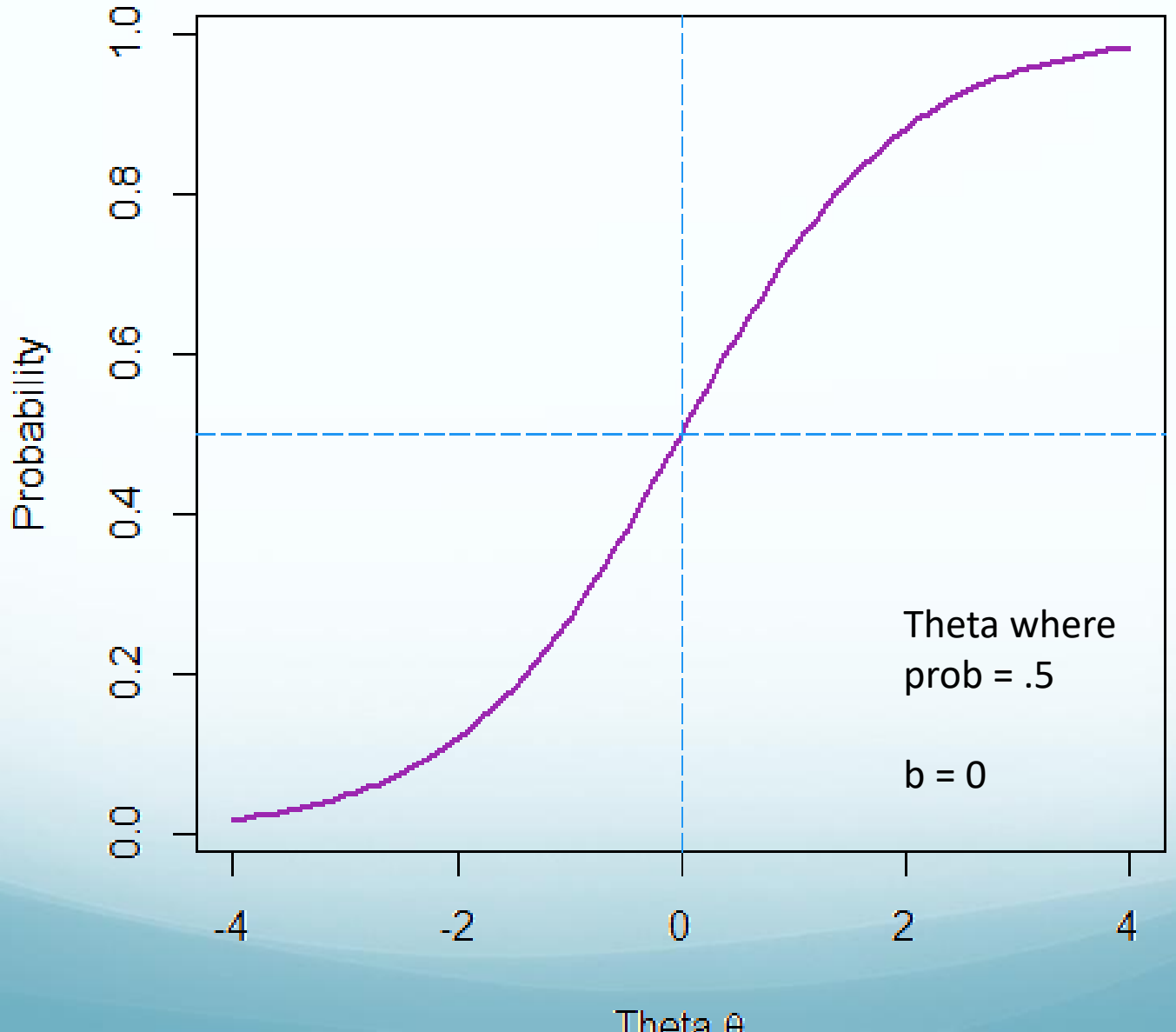
If an item has a high θ value, what does that mean in terms of difficulty?

If a person has a low θ value, what does that mean in terms of ability?

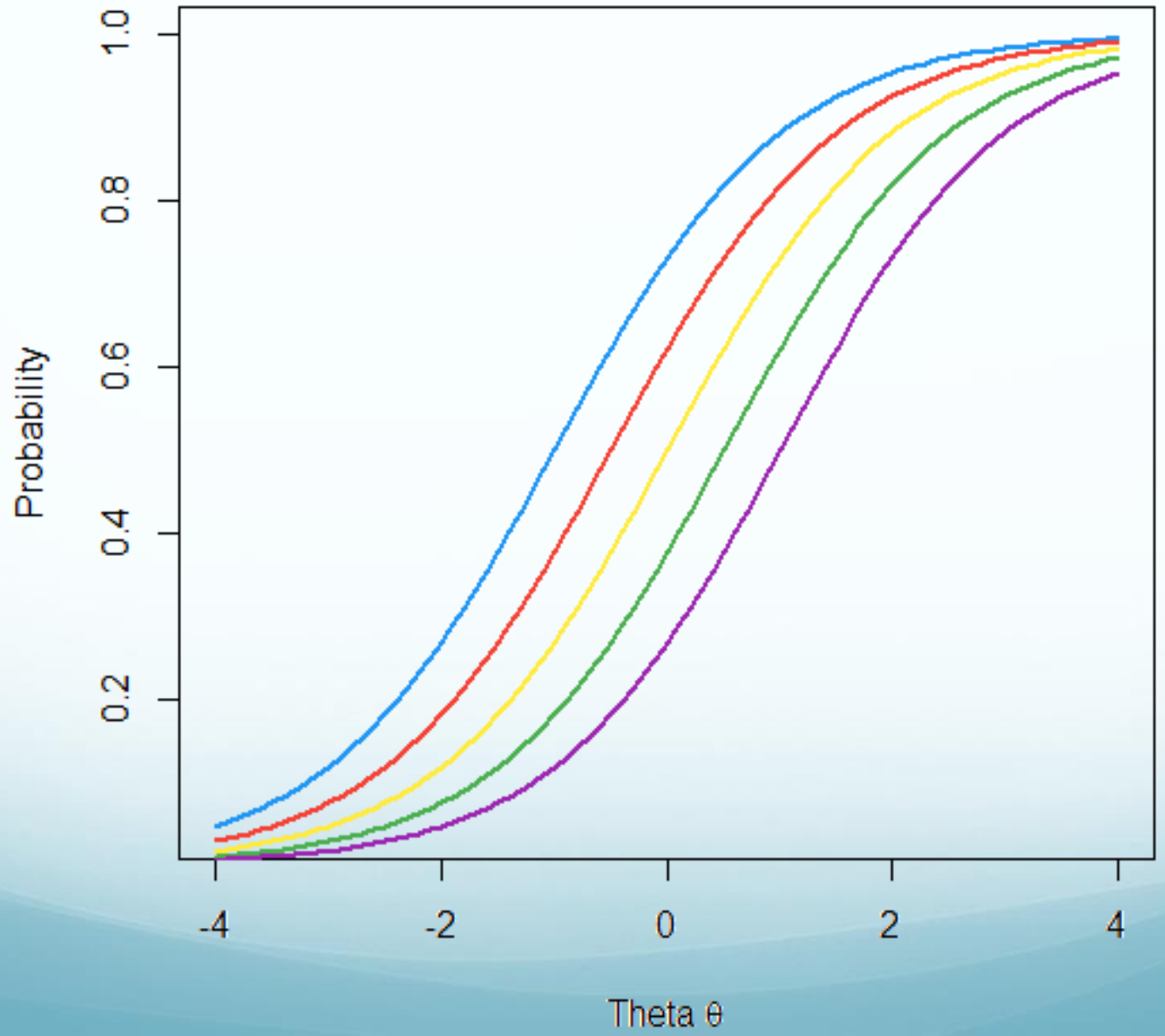
Item response function



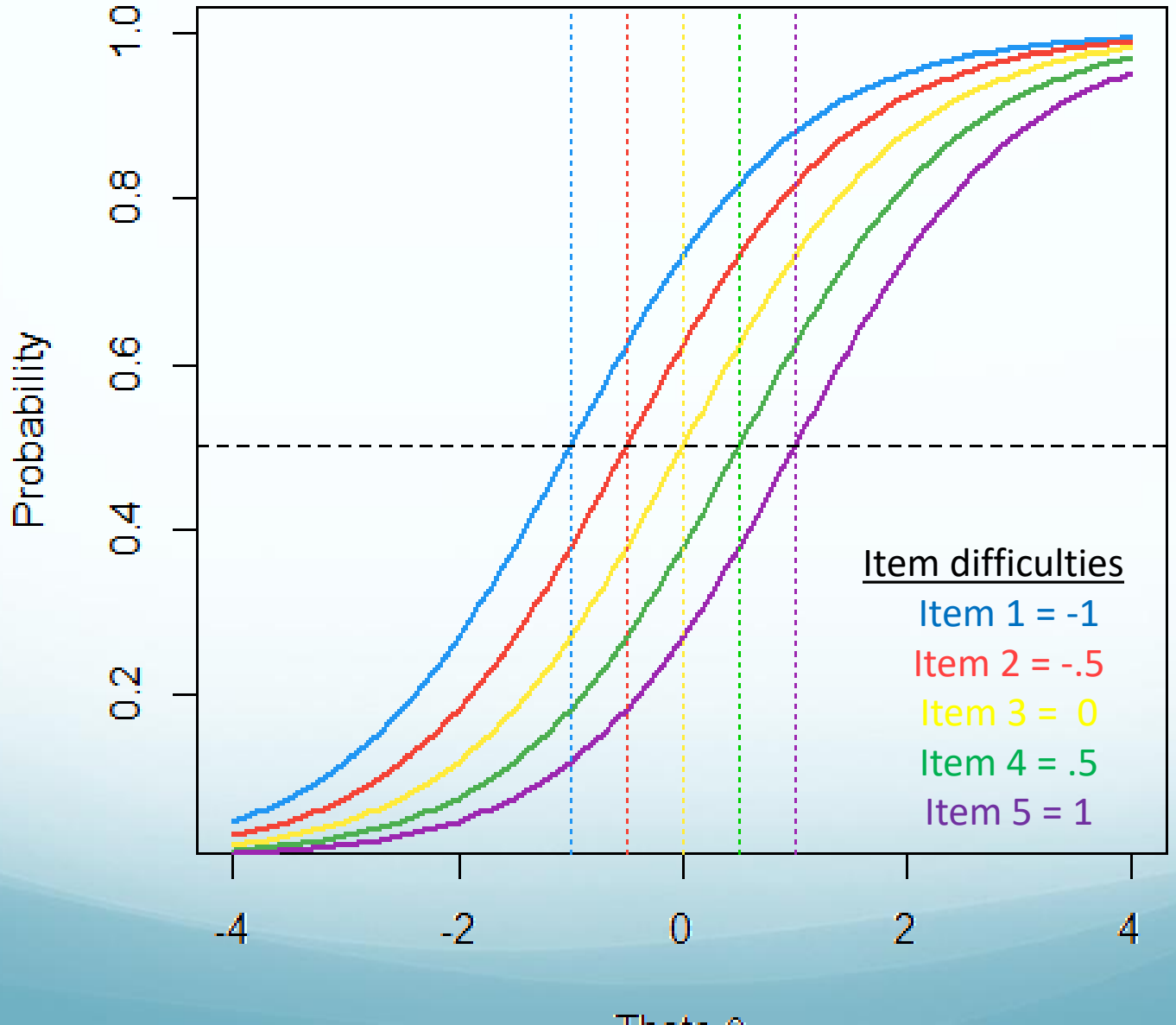
Item ‘difficulty’ (b) parameter (1PL)



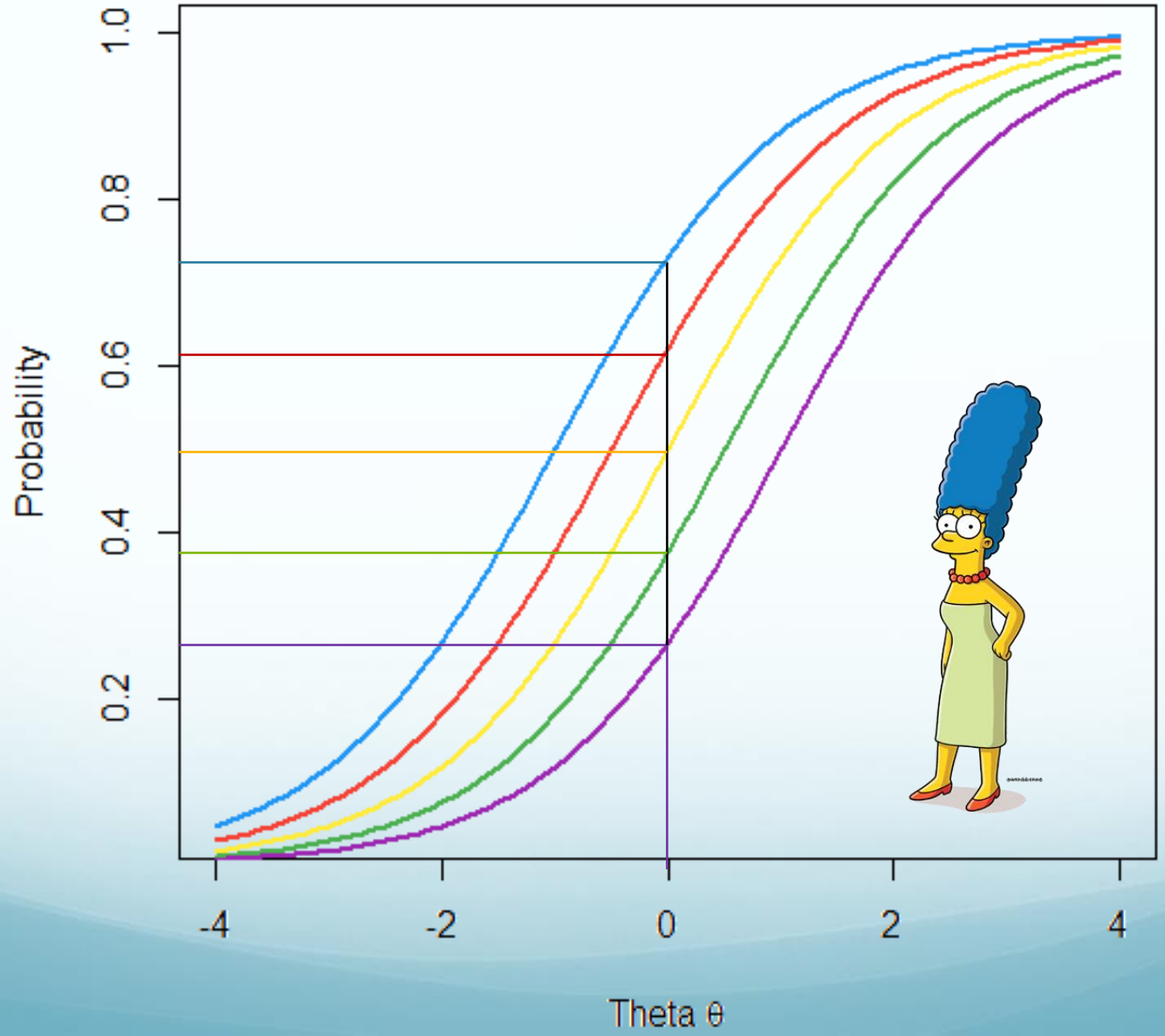
Item ‘difficulty’ (b) parameter (1PL)



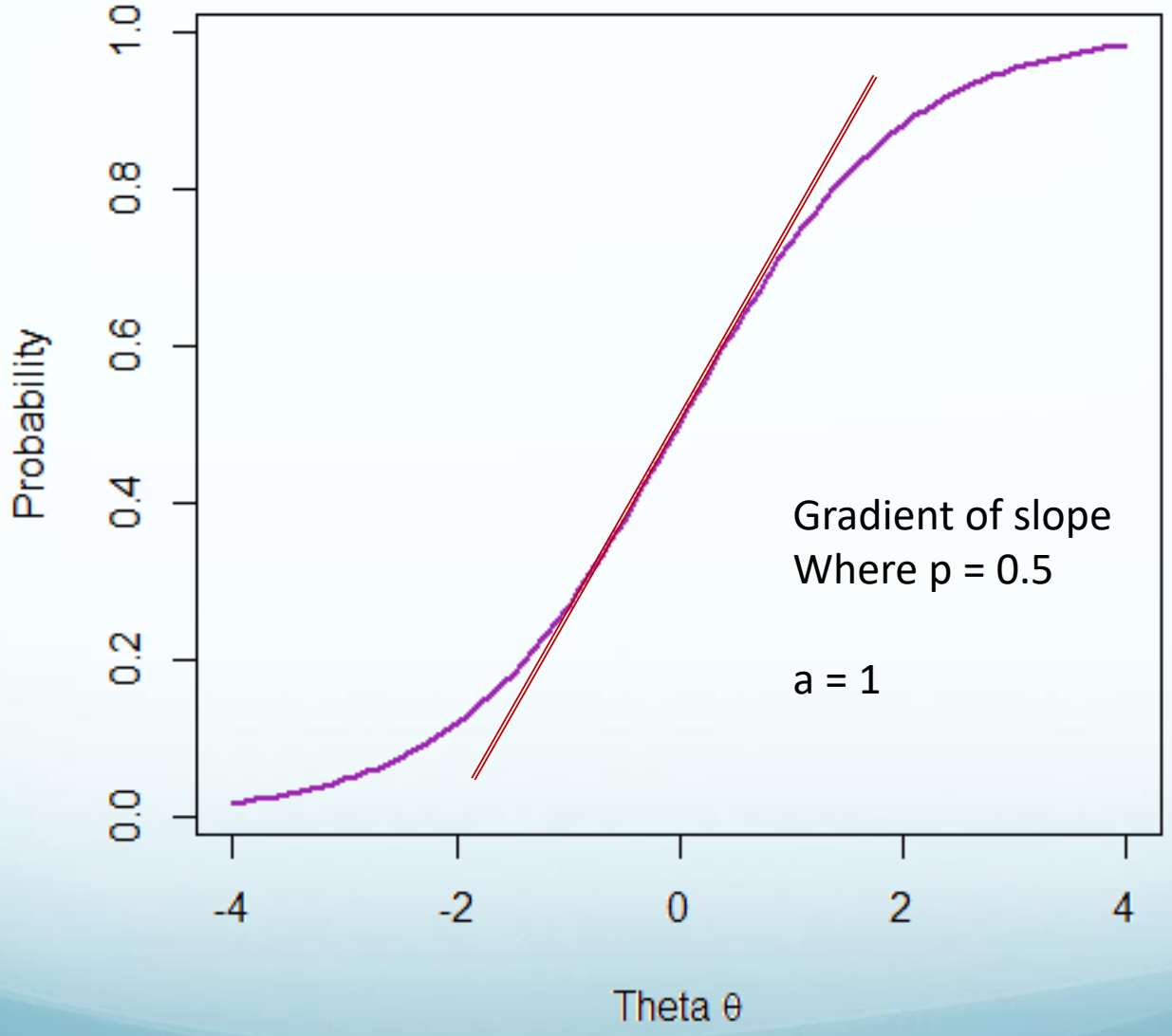
Item ‘difficulty’ (b) parameter (1PL)



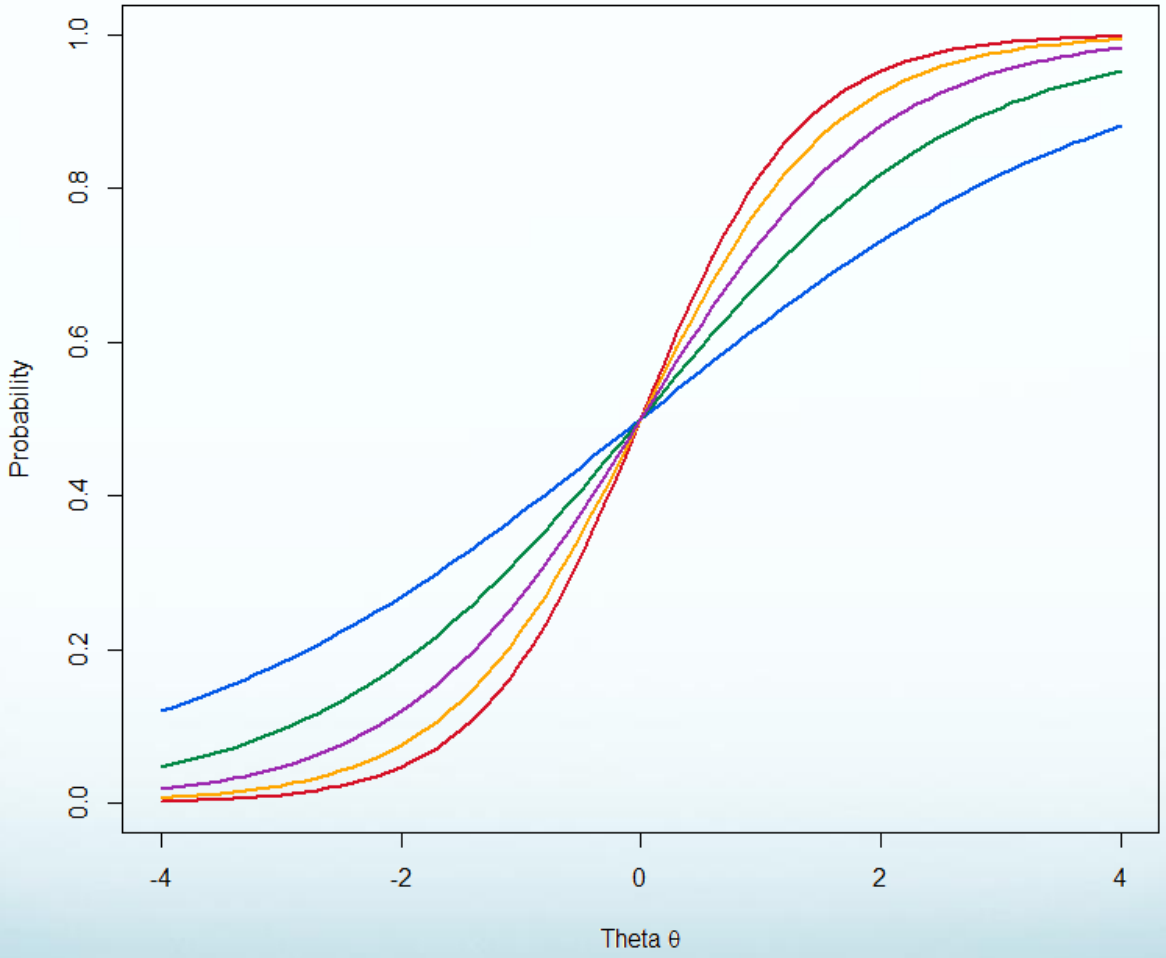
Item ‘difficulty’ (b) parameter (1PL)



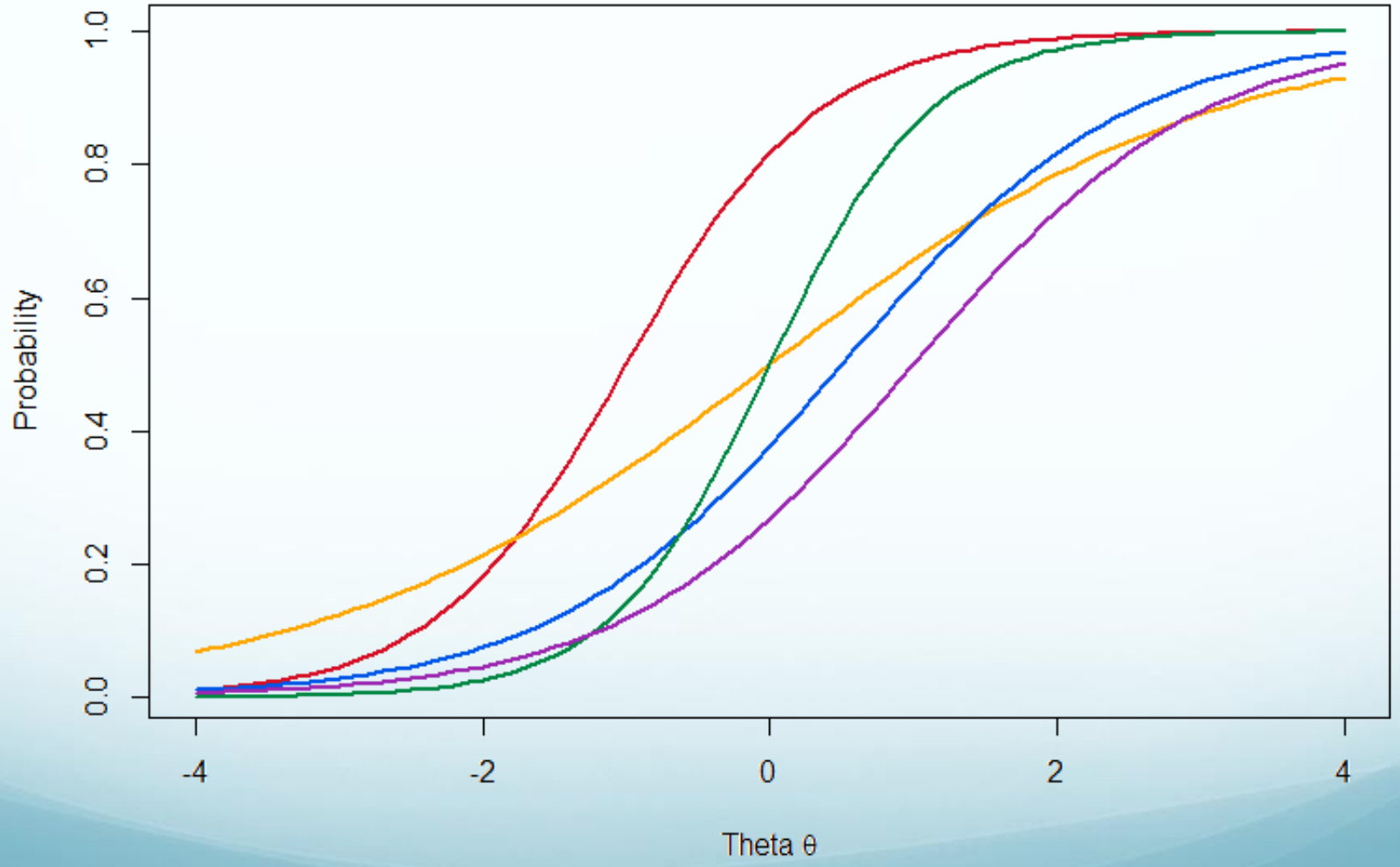
Item ‘discrimination’ (a) parameter



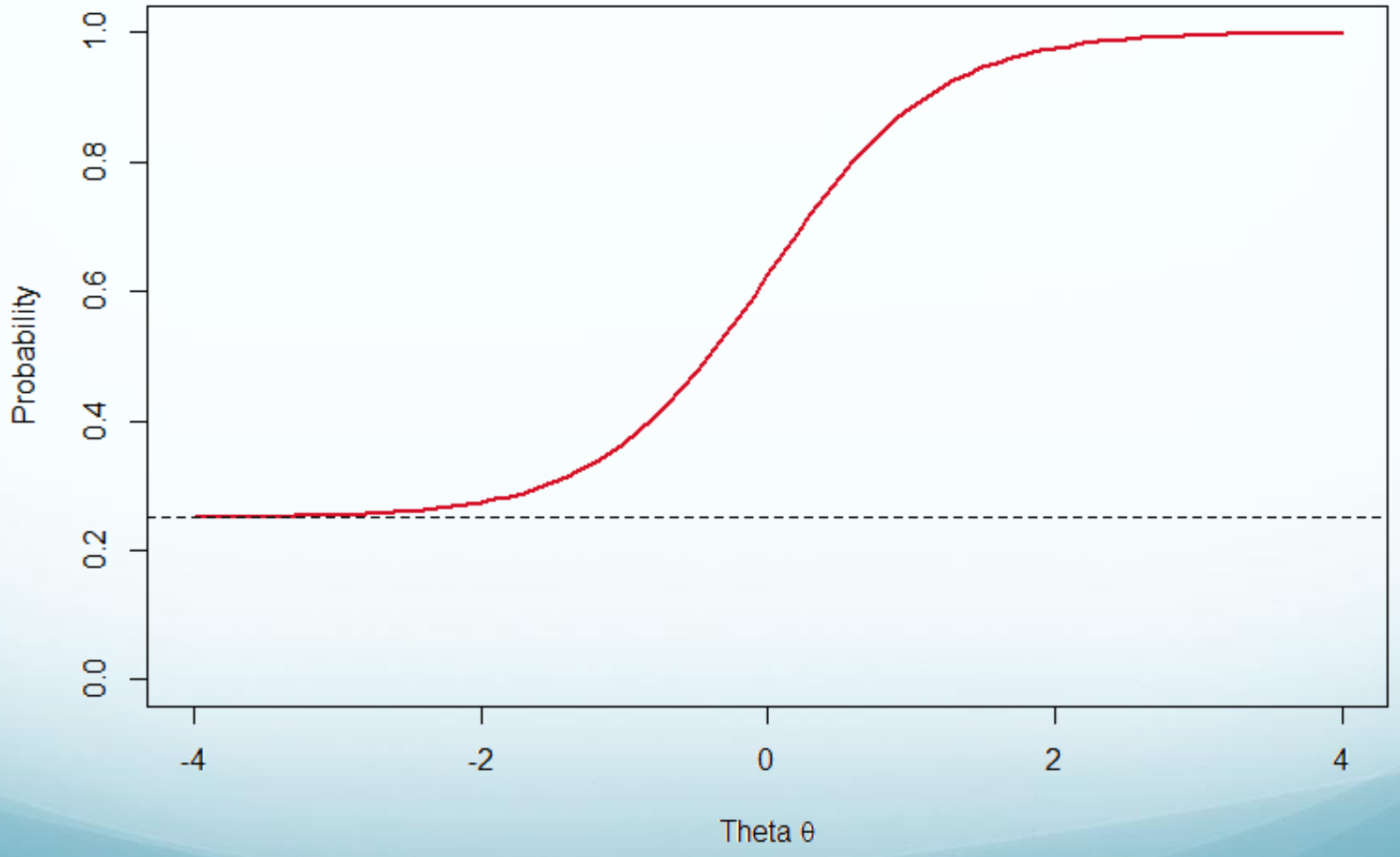
Item ‘discrimination’ (a) parameter



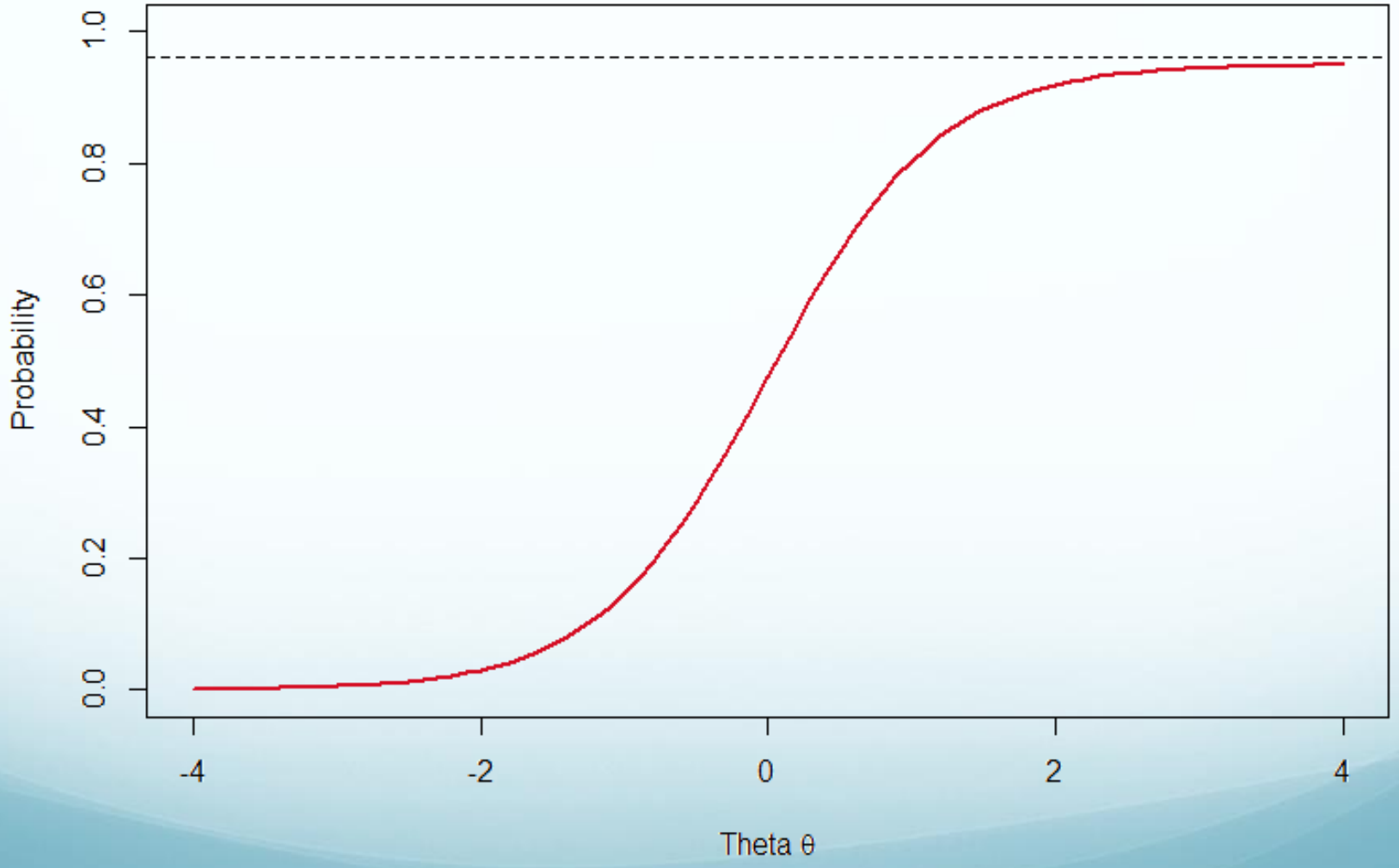
Item ‘discrimination’ (a) parameter
And item ‘difficulty’ (b) parameter



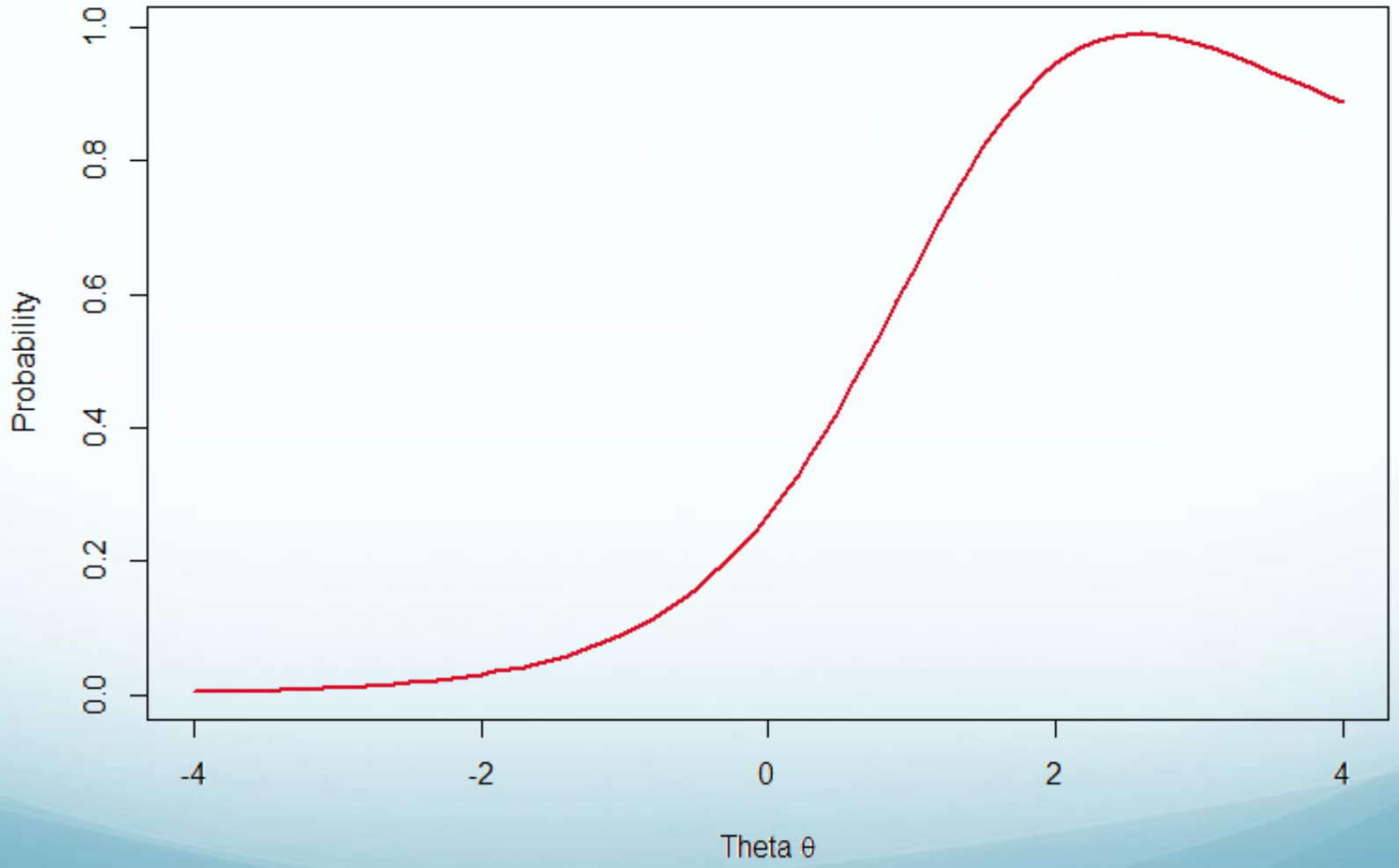
Guessing (c) parameter



Inattention (d) parameter



Unfolding (e) parameter



Different IRT Models

Non-parametric
- Mokken

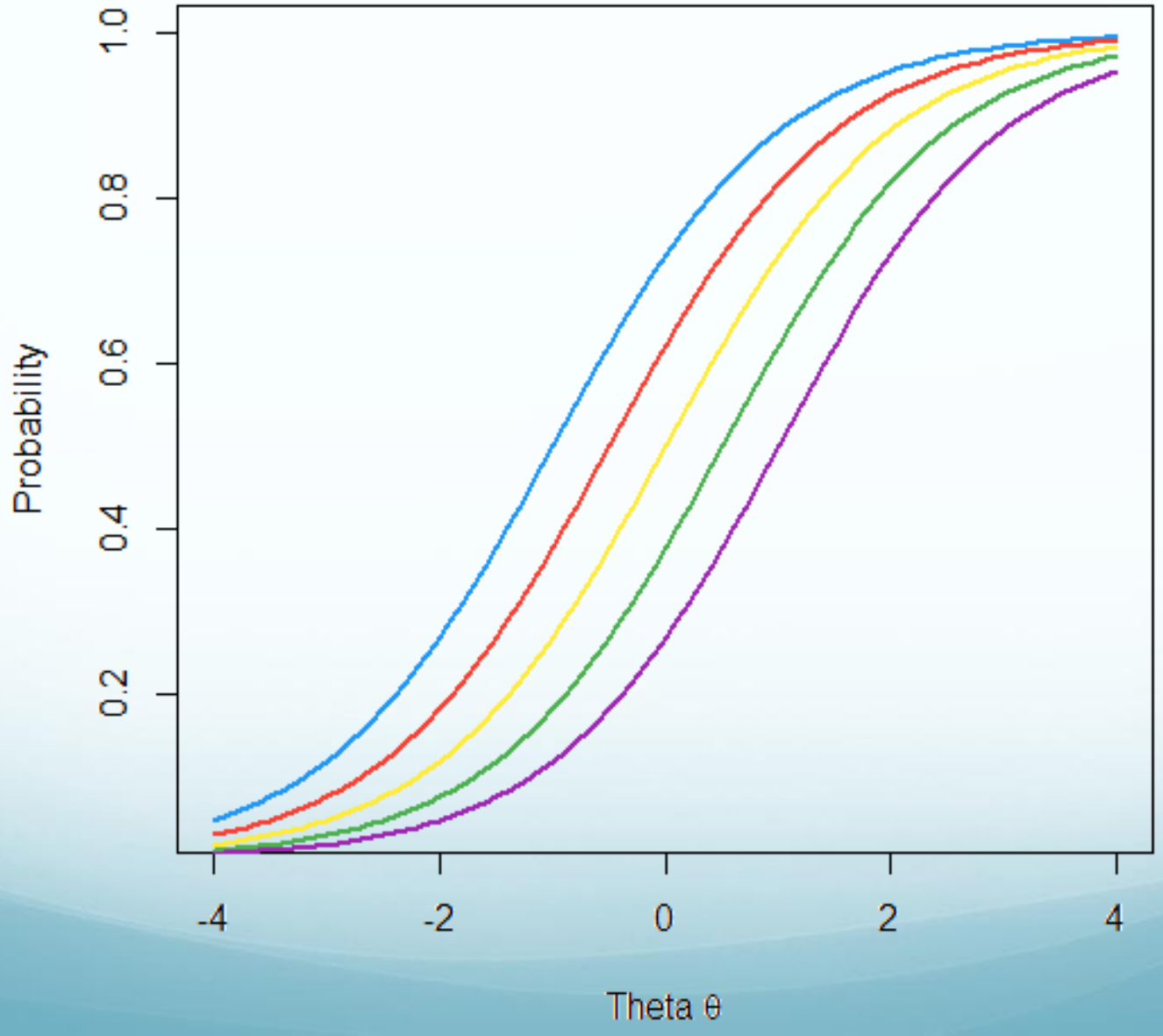
1-parameter (1pl) models (only item difficulty changes)
- Rasch model (for dichotomous data)
- Partial Credit Model (for polytomous data)
- Rating Scale Model (for polytomous data)

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

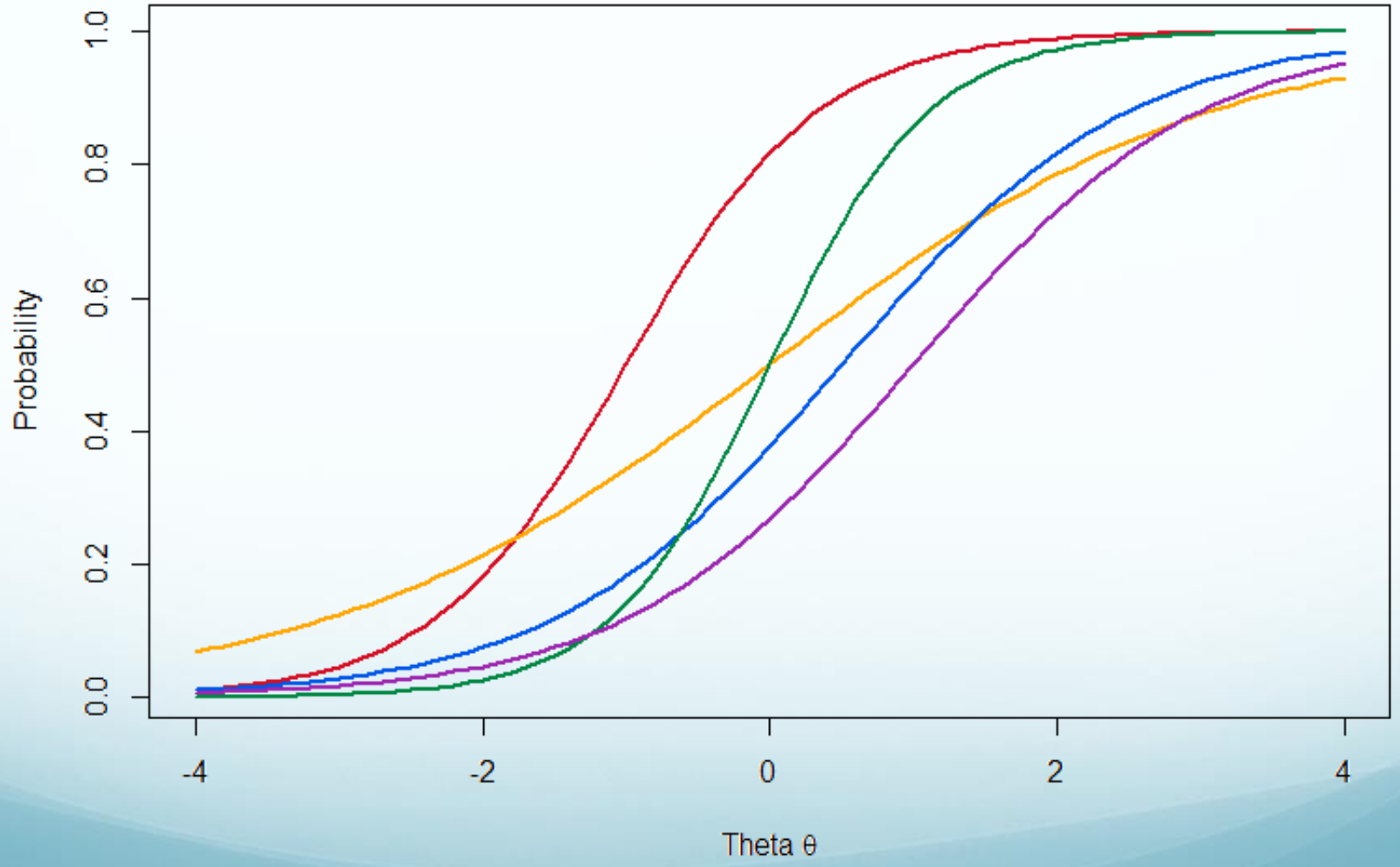
3+ parameter (3pl) models (with a guessing/inattention/unfolding parameter)
- Three-parameter logistic model (etc..)

Multidimensional
- Compensatory
- Bi-factor
- …

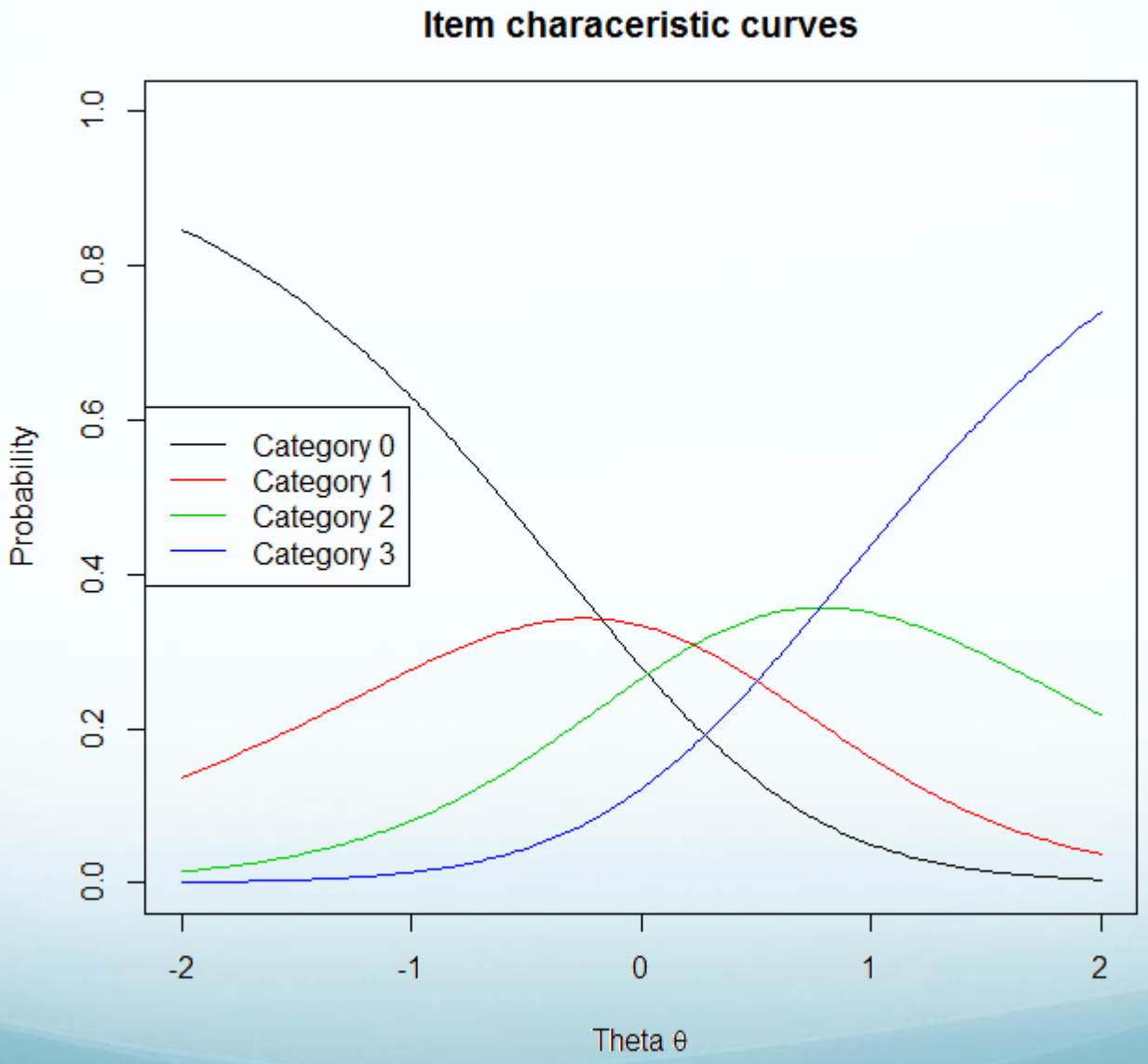
One parameter / Rasch model



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

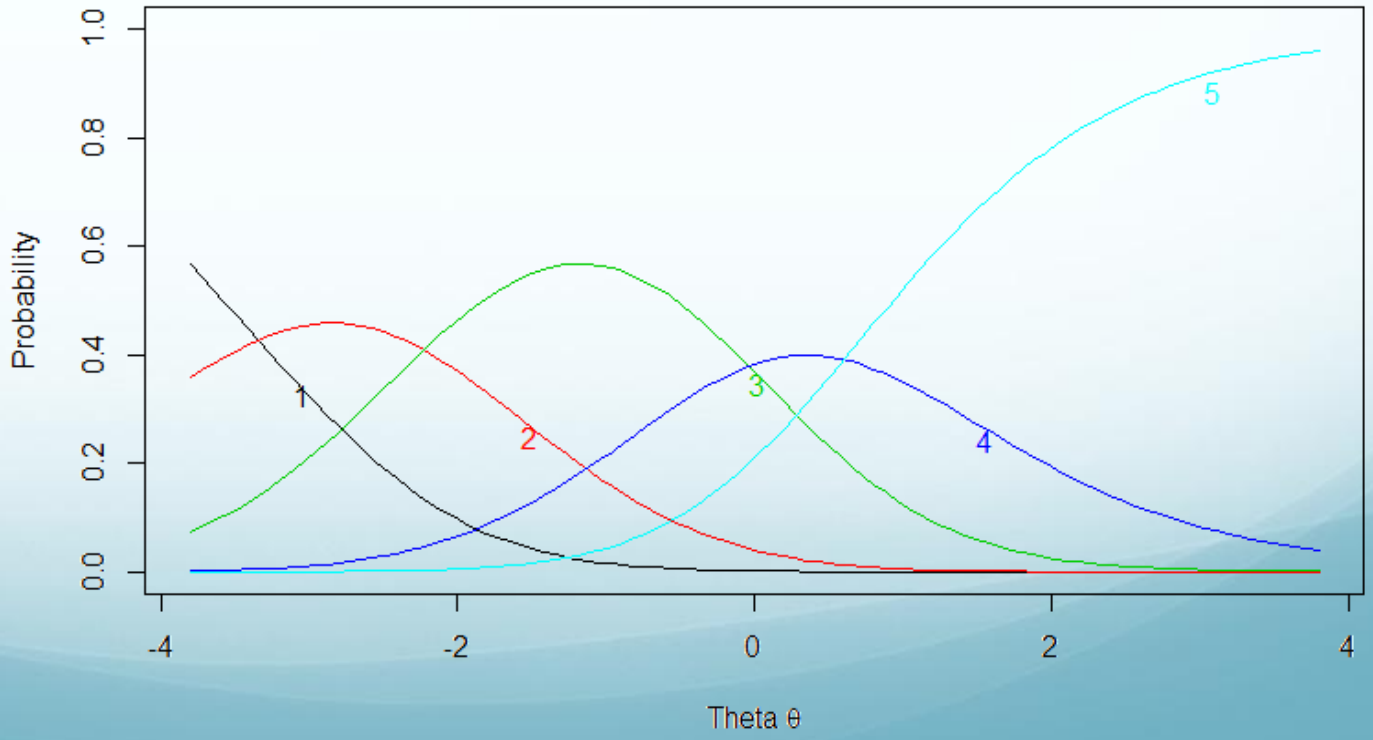
- Or we can investigate *why* the model is not fitting, using some known criteria that are liable to cause misfit

  - 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

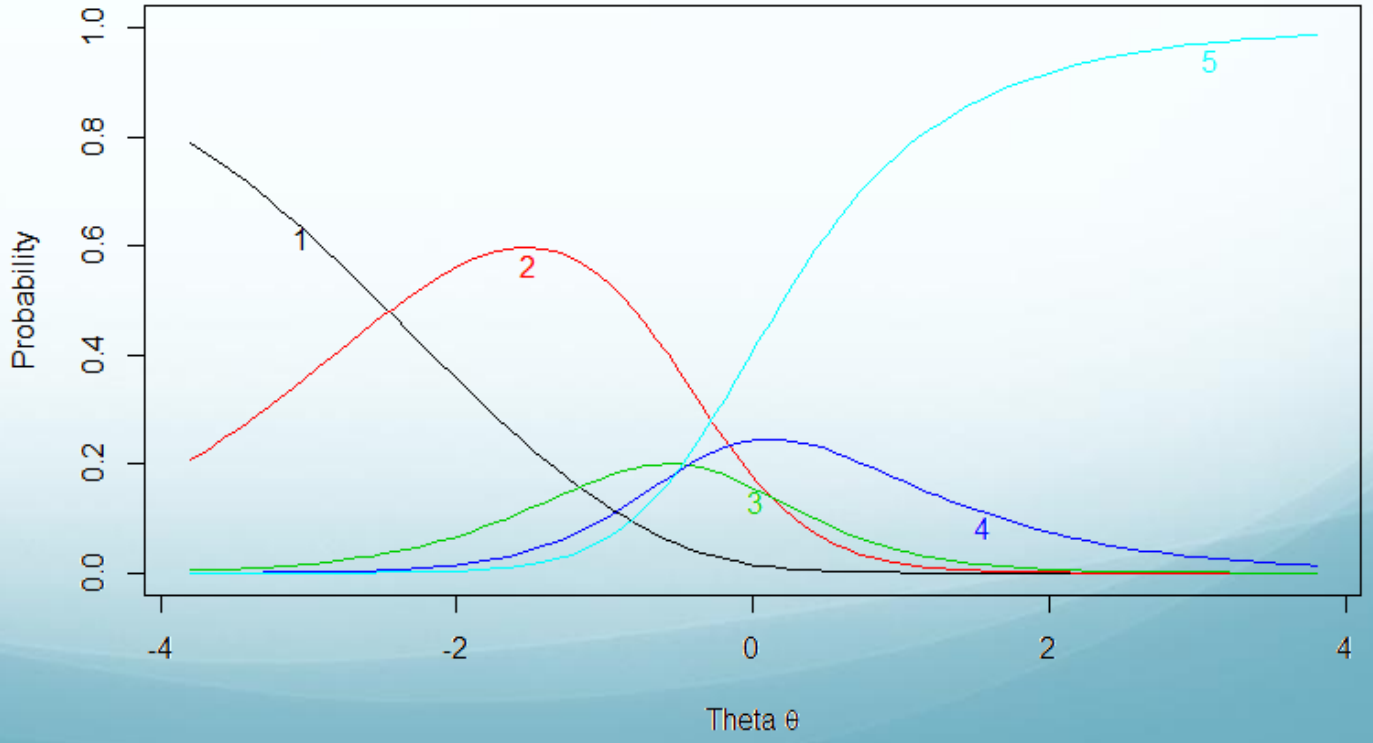
Category threshold ordering

How much do any difficulties in mobility bother you?				
Not at all	A little	A moderate amount	Very much	An extreme amount
1	2	3	4	5



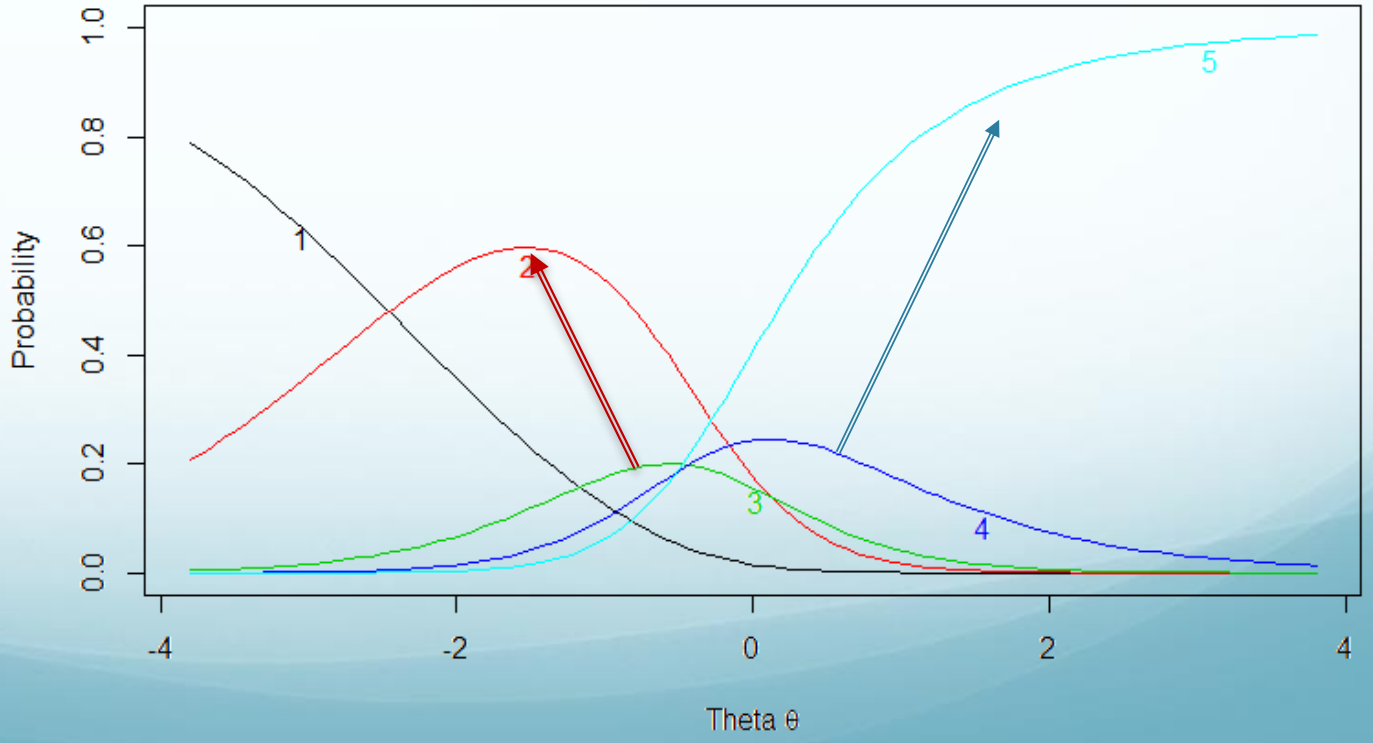
Category threshold ordering

How much do any difficulties in mobility bother you?				
Not at all	A little	A moderate amount	Very much	An extreme amount
1	2	3	4	5



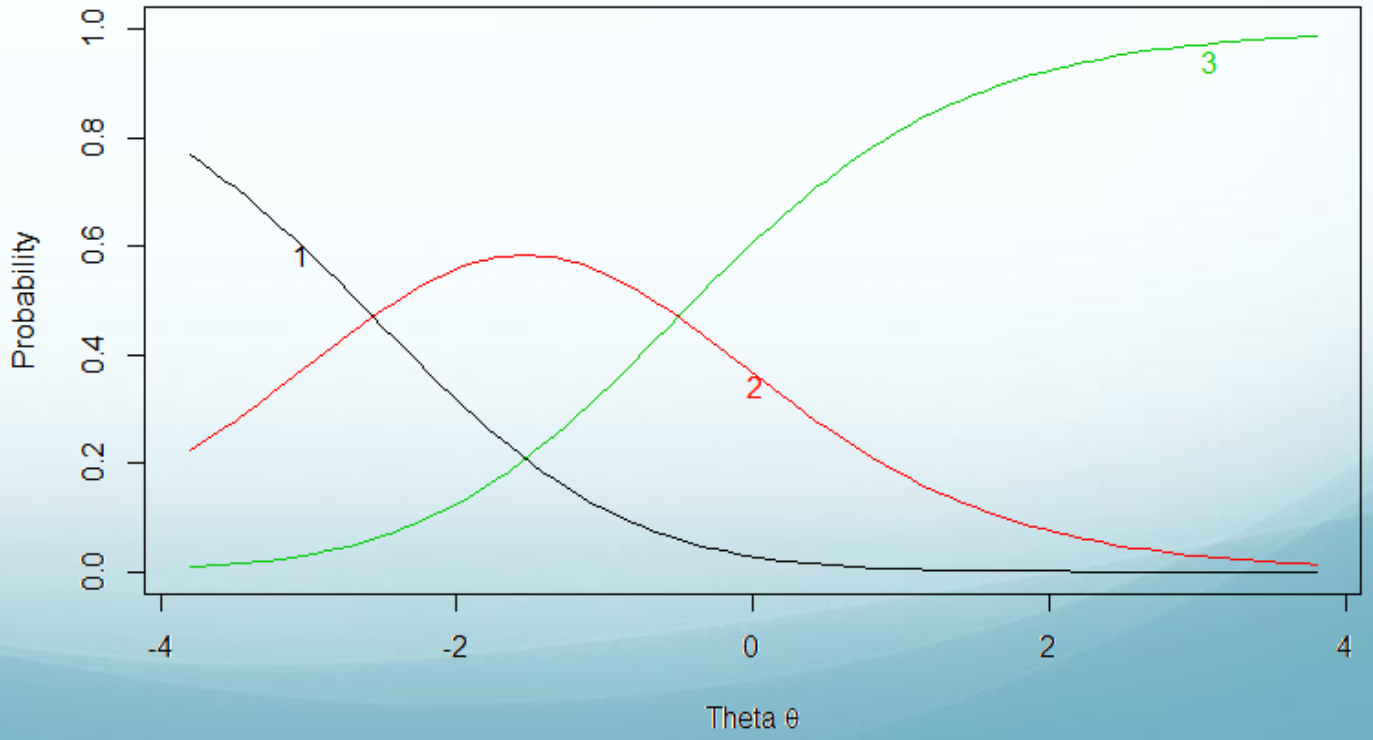
Category threshold ordering

How much do any difficulties in mobility bother you?				
Not at all	A little	A moderate amount	Very much	An extreme amount
1	2	3	4	5



Category threshold ordering

How much do any difficulties in mobility bother you?				
Not at all	A little	A moderate amount	Very much	An extreme amount
1	2	2	3	3



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 –

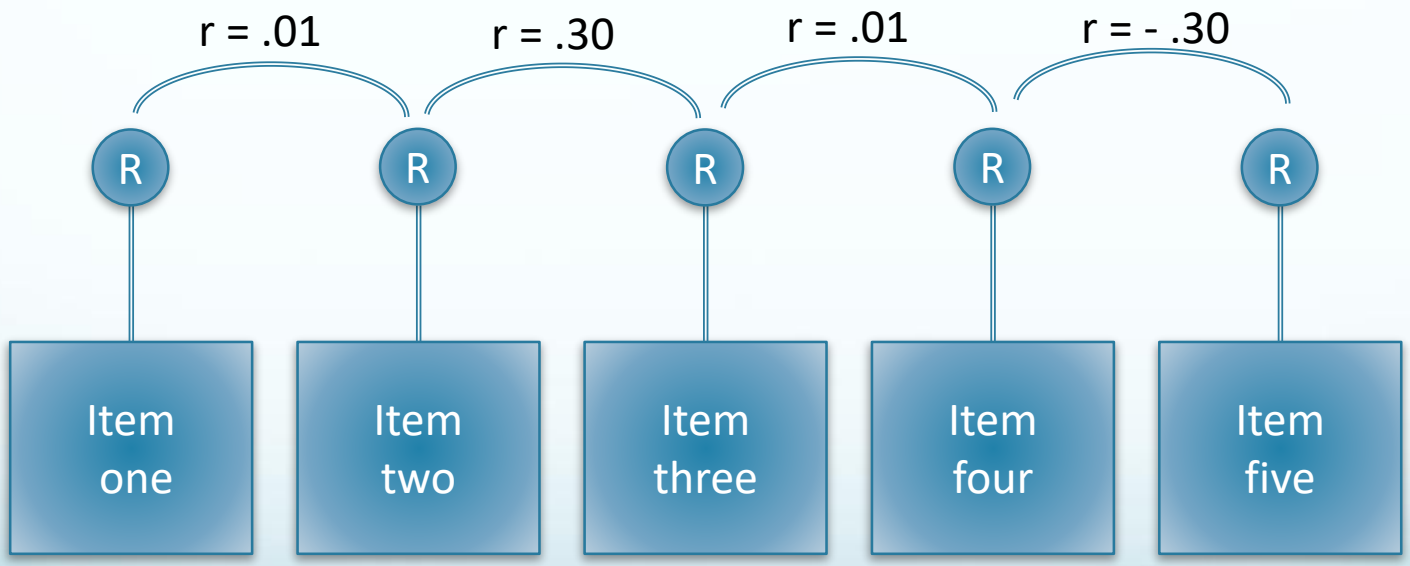
- Walking more than a kilometer
- Walking more than half a kilometer
- Walking more than 100m

- Or –

“I was happy”
“I enjoyed life ”

Local dependency

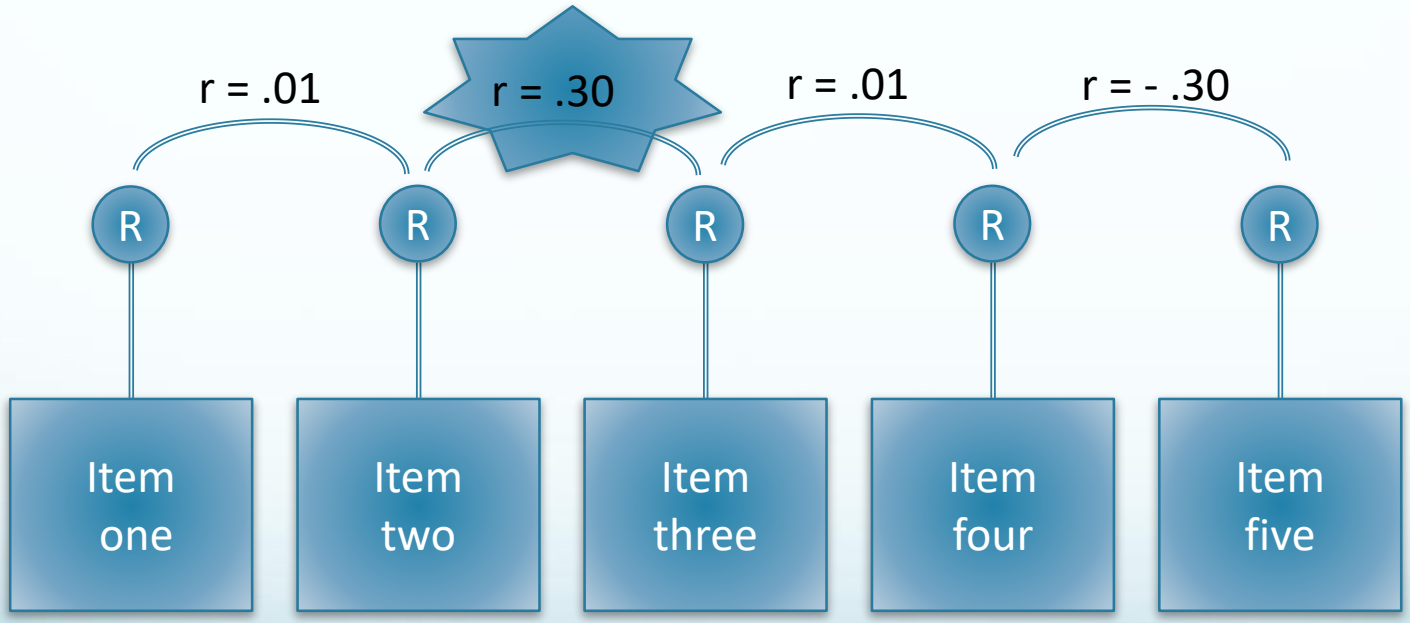
- Yen's Q3
- Correlation between the item residuals
- Cut of off +.2 is indicative of local dependency



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

Local dependency

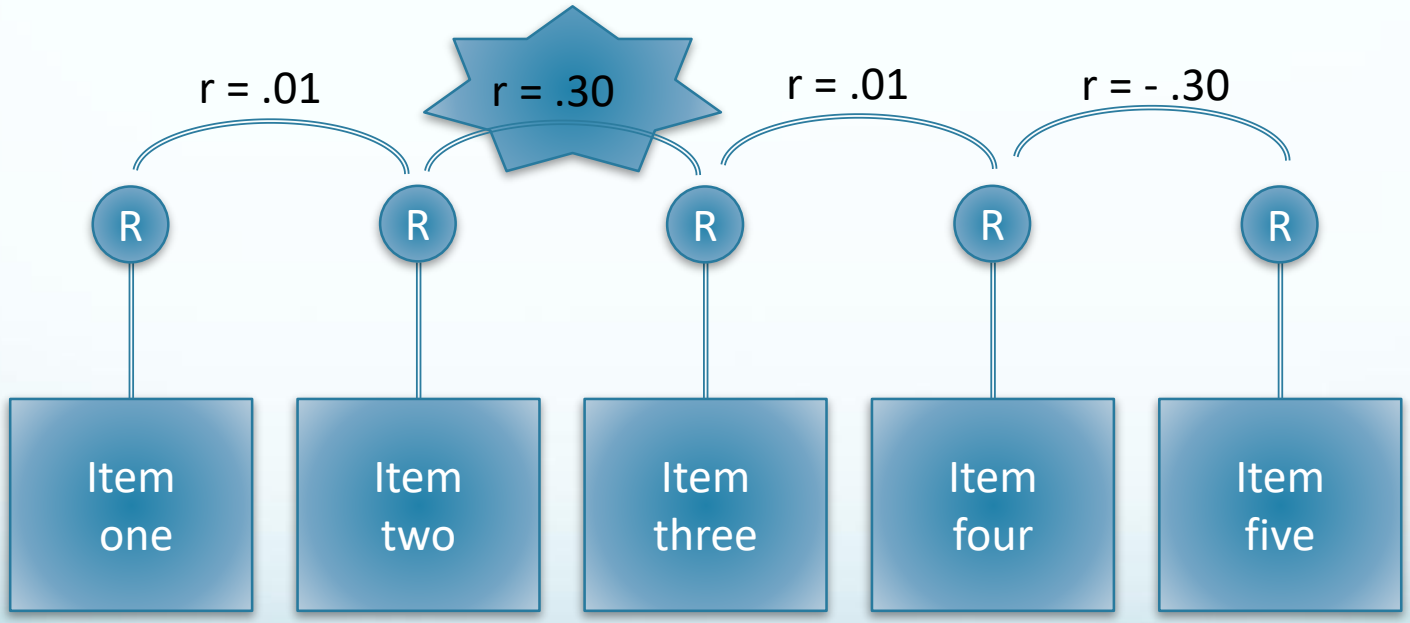
- Yen's Q3
- Correlation between the item residuals
- Cut of off +.2 is indicative of local dependency



Item two = "Are you able to work? ", Item three = "How would you rate your ability to work? "

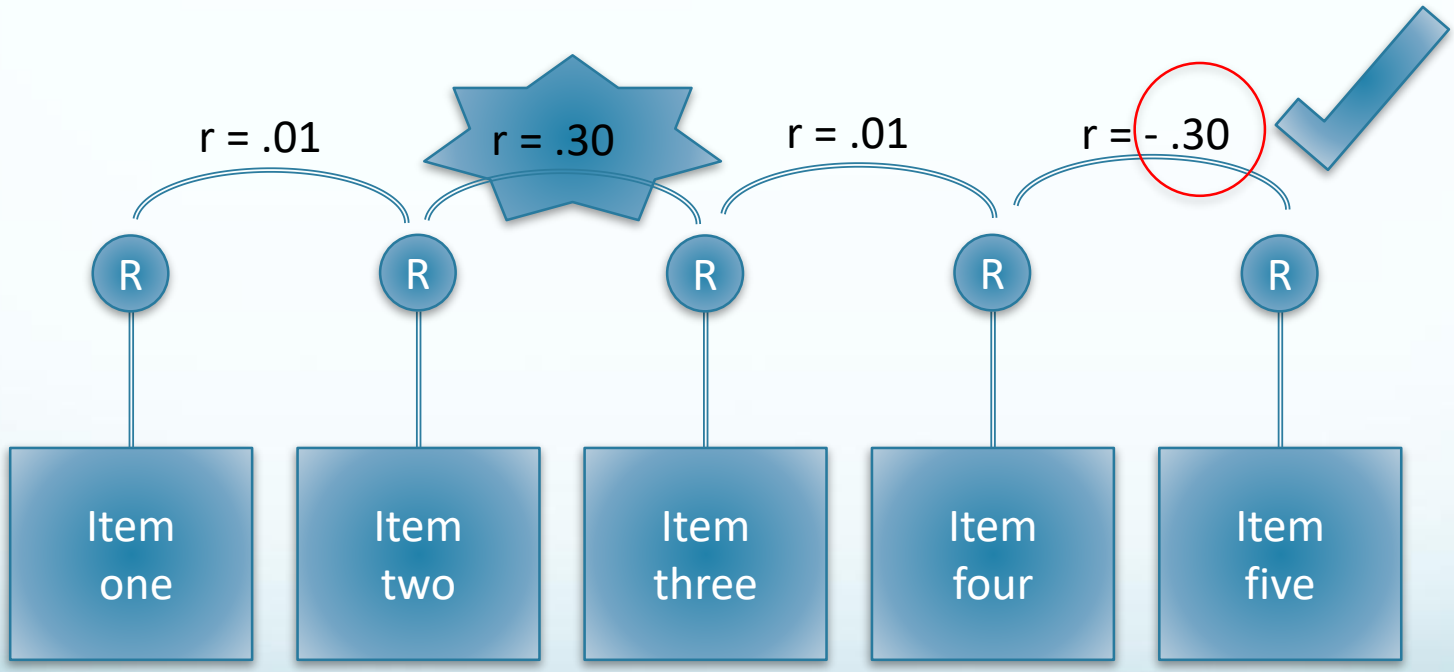
Local dependency

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



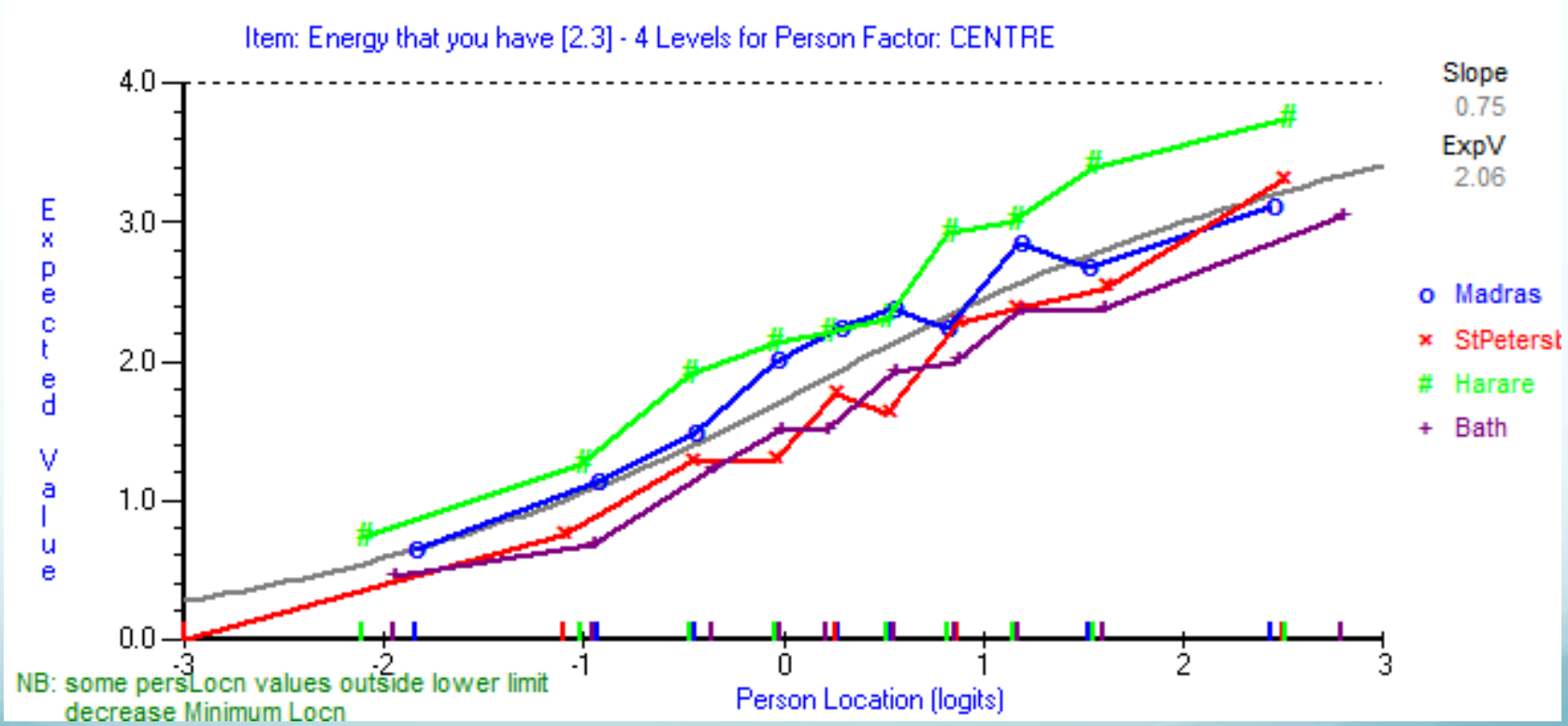
Local dependency

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



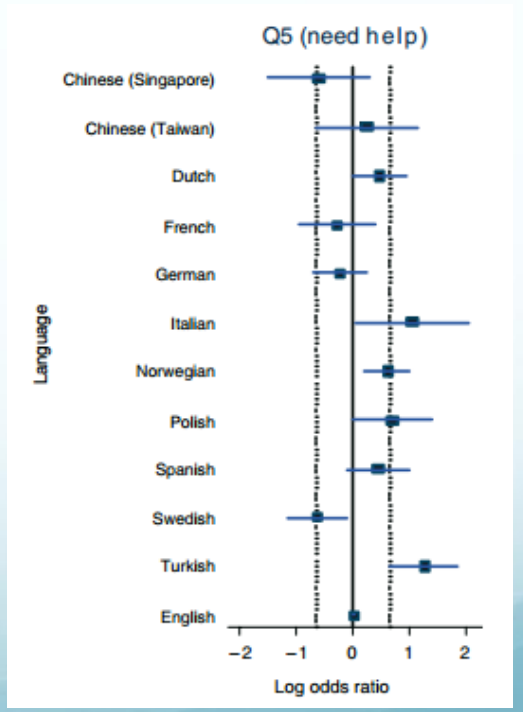
Differential item functioning

- 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

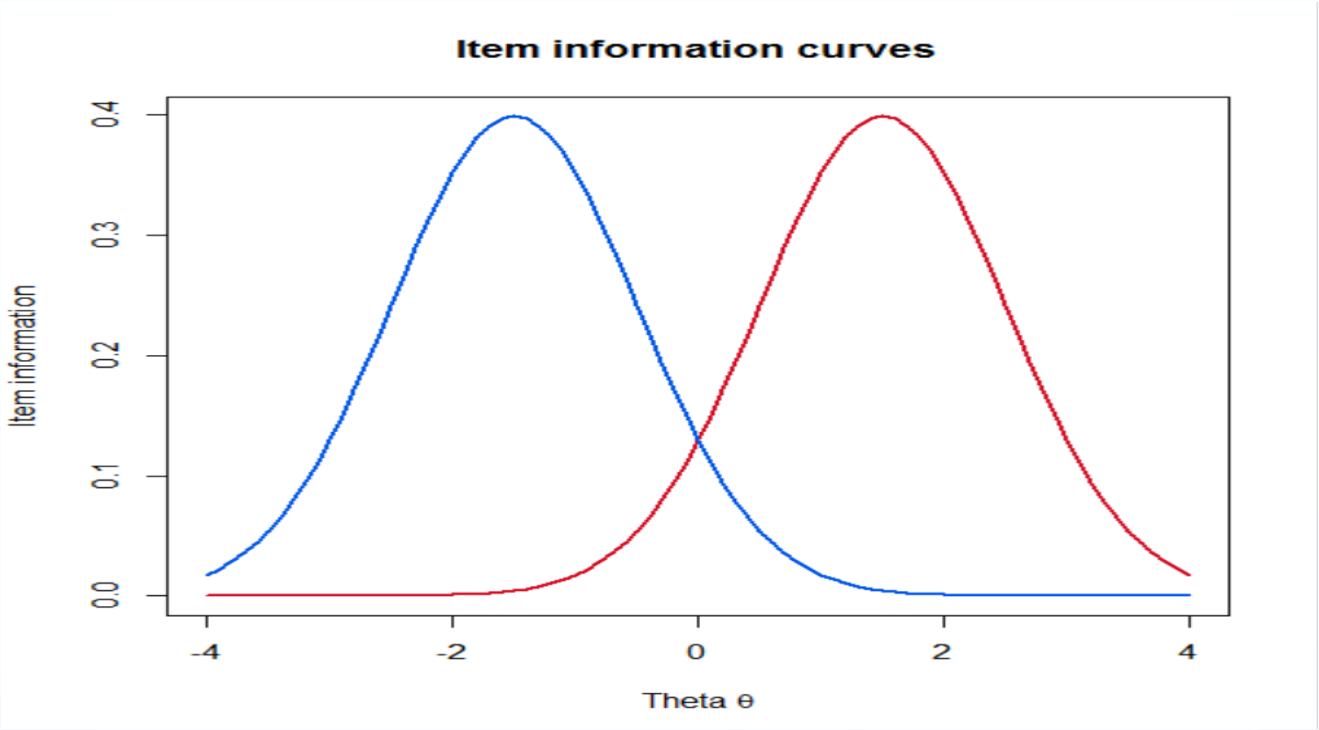
Reliability

- 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!

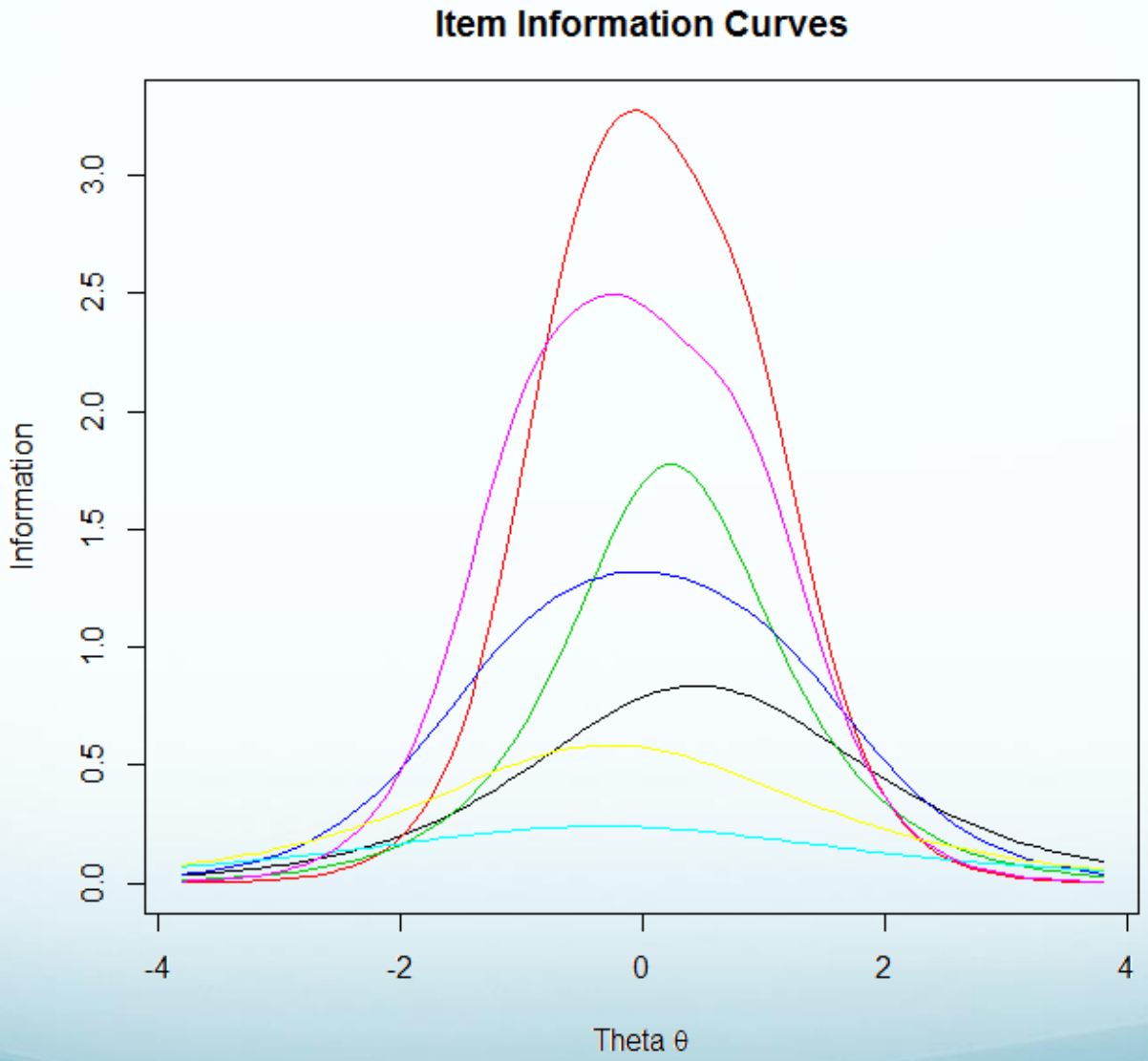
Item information, standard error and reliability are all related

Individual item information

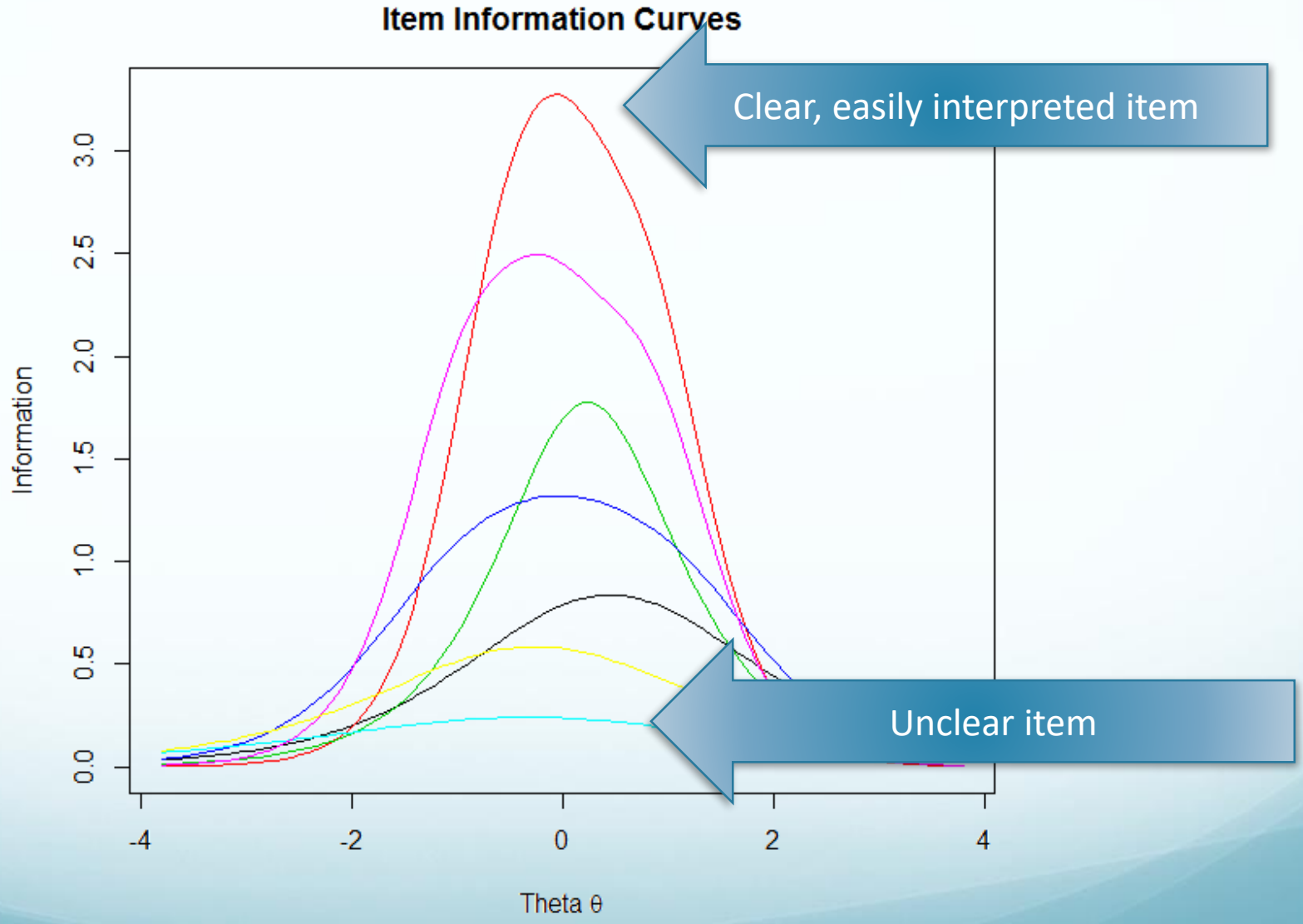


- Blue item "Sometimes I feel a bit sad"
- Red item "I often feel suicidal"

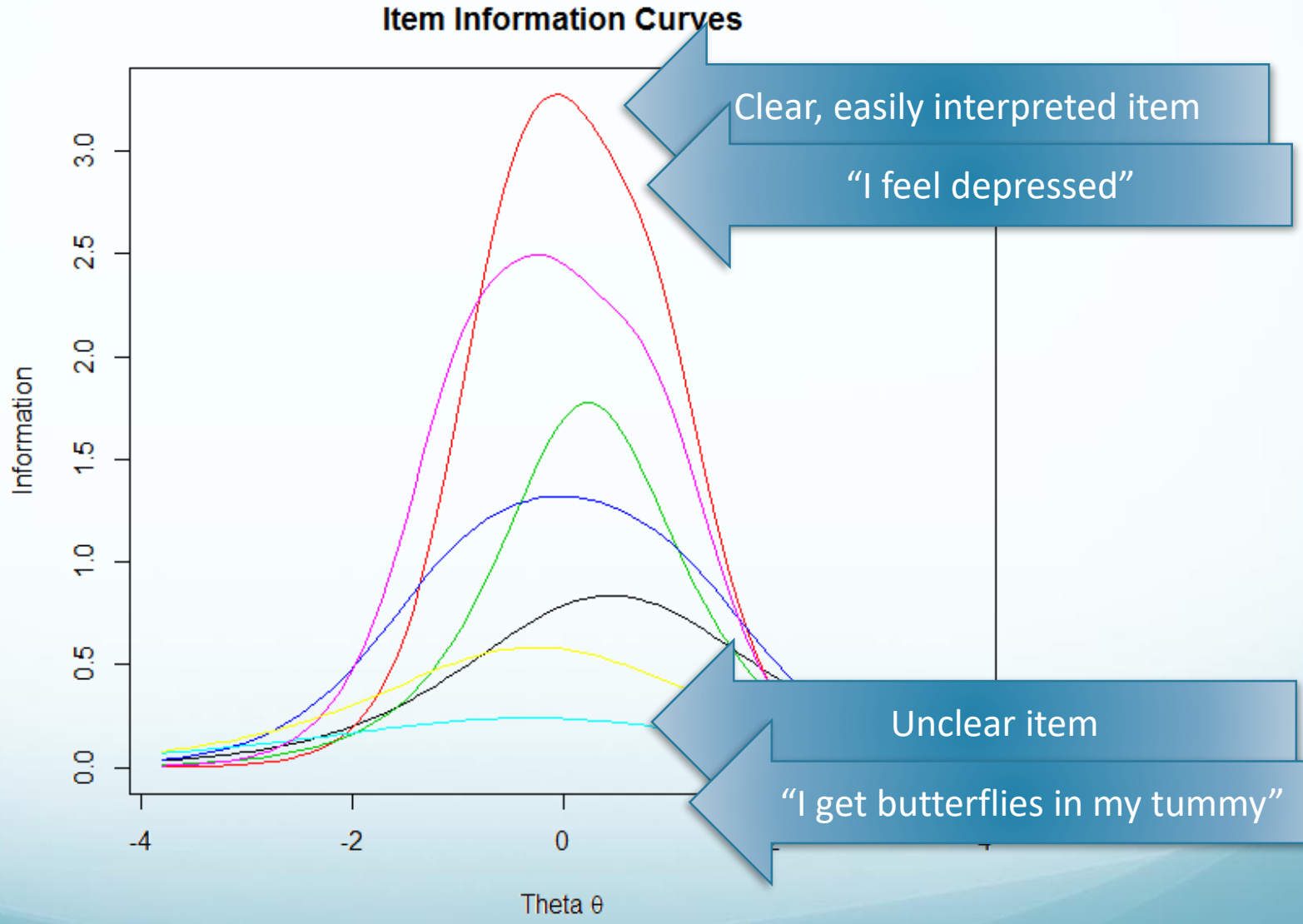
Individual item information



Individual item information

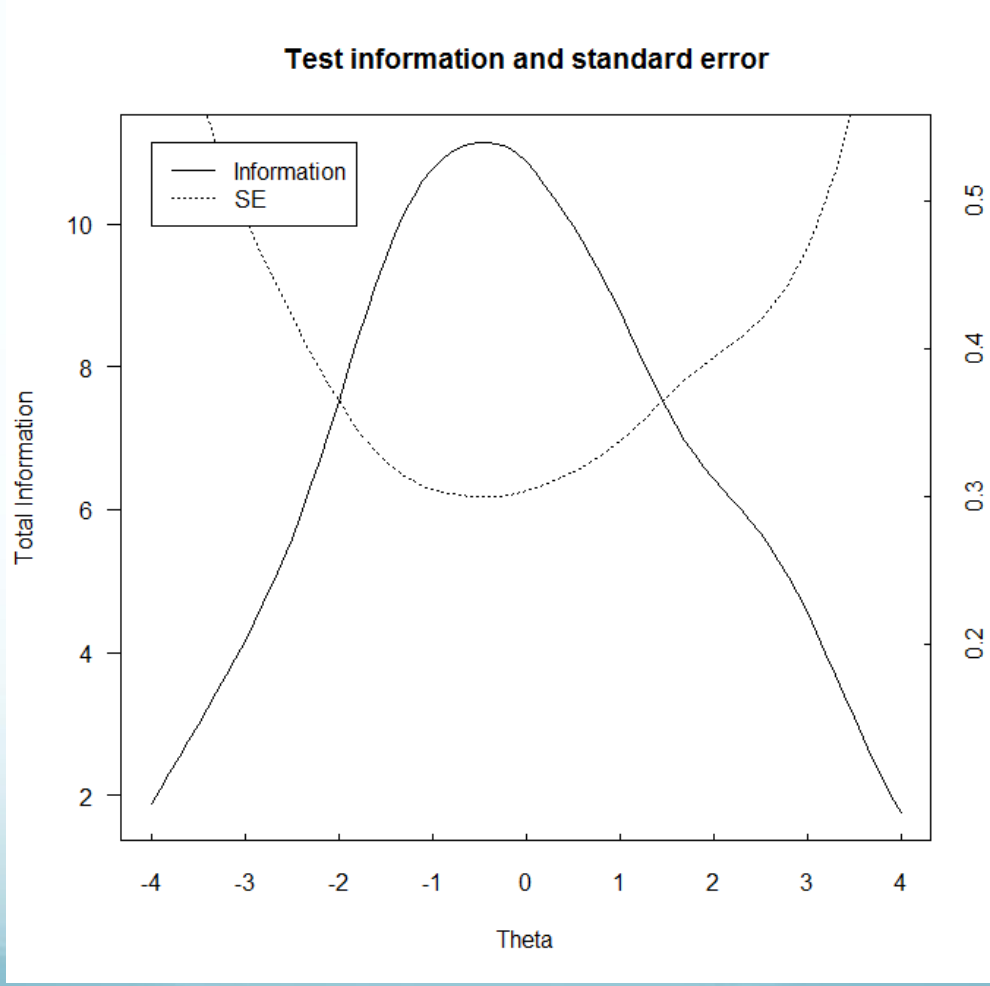


Individual item information



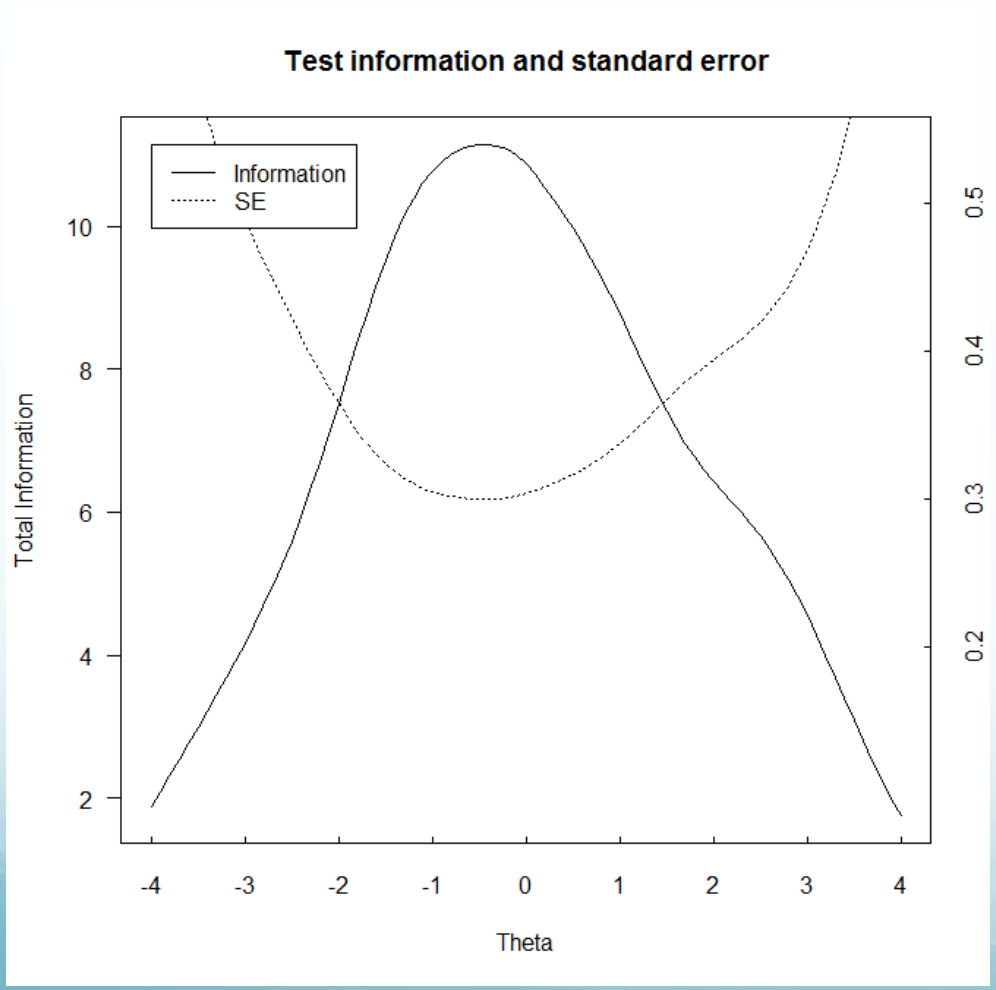
Test information

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



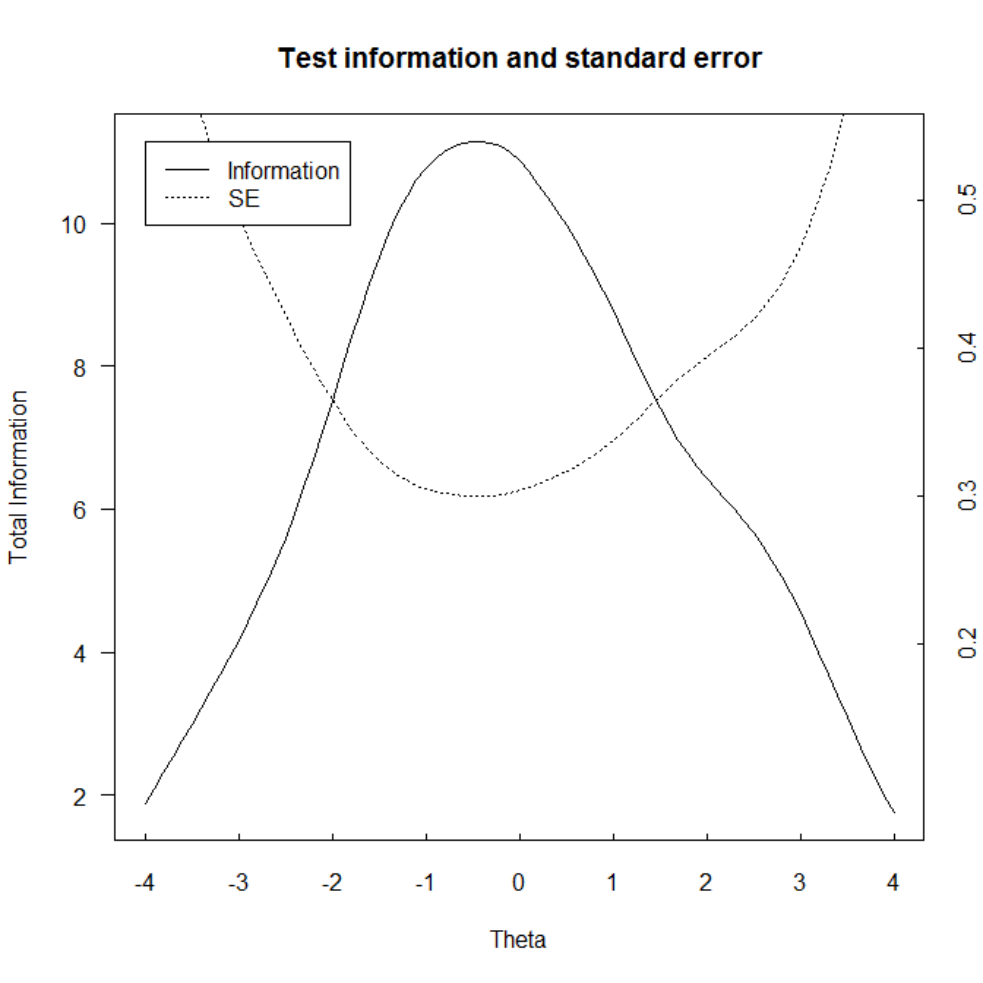
Information and Standard Error

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



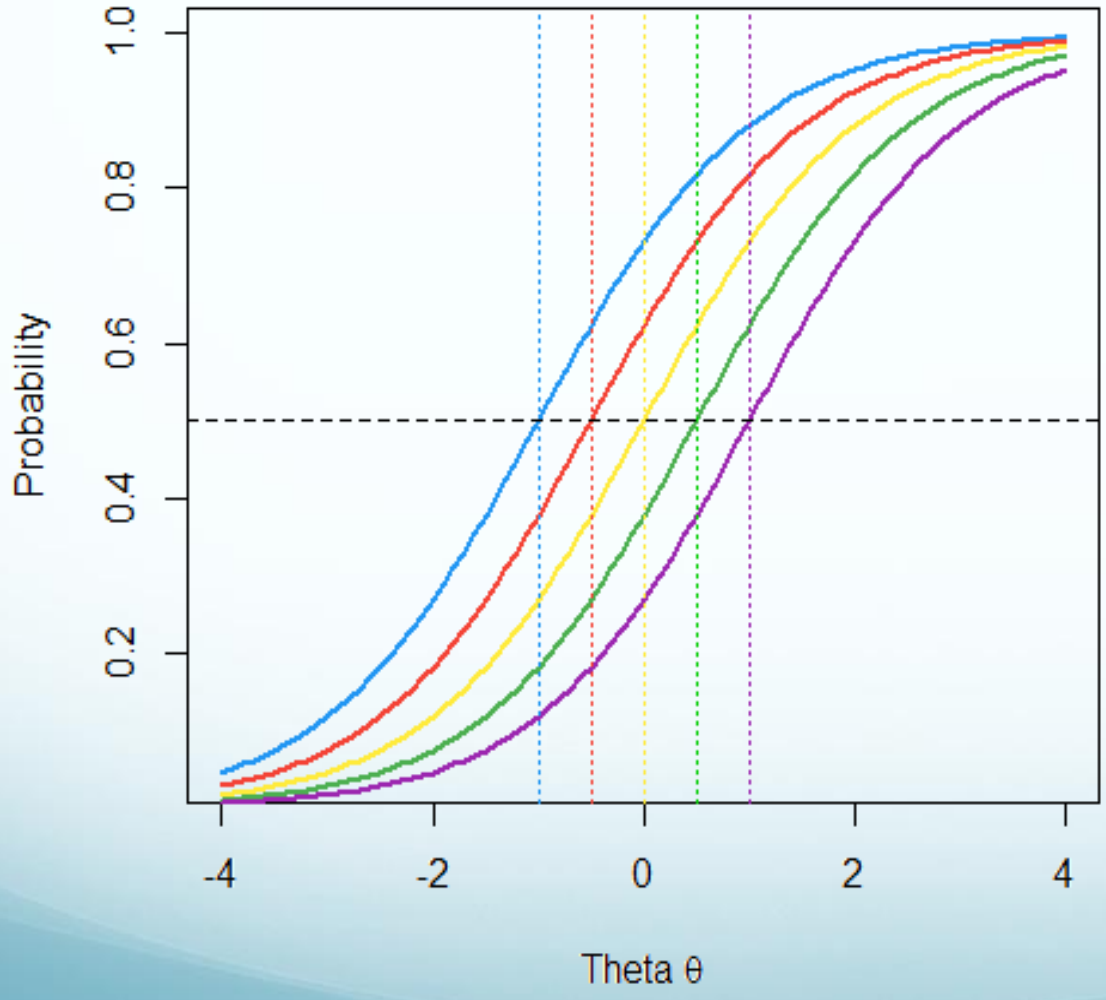
- Error of measurement inversely related to information
- Standard error (SE) is an estimate of measurement precision at a given theta
- Reliability can be calculated from SE and information

Information and Reliability



$$Reliability\ (\theta) = 1 - \left(\frac{1}{\sqrt{I(\theta)}}\right)\text{^}2$$

IRT parameters (1pl)

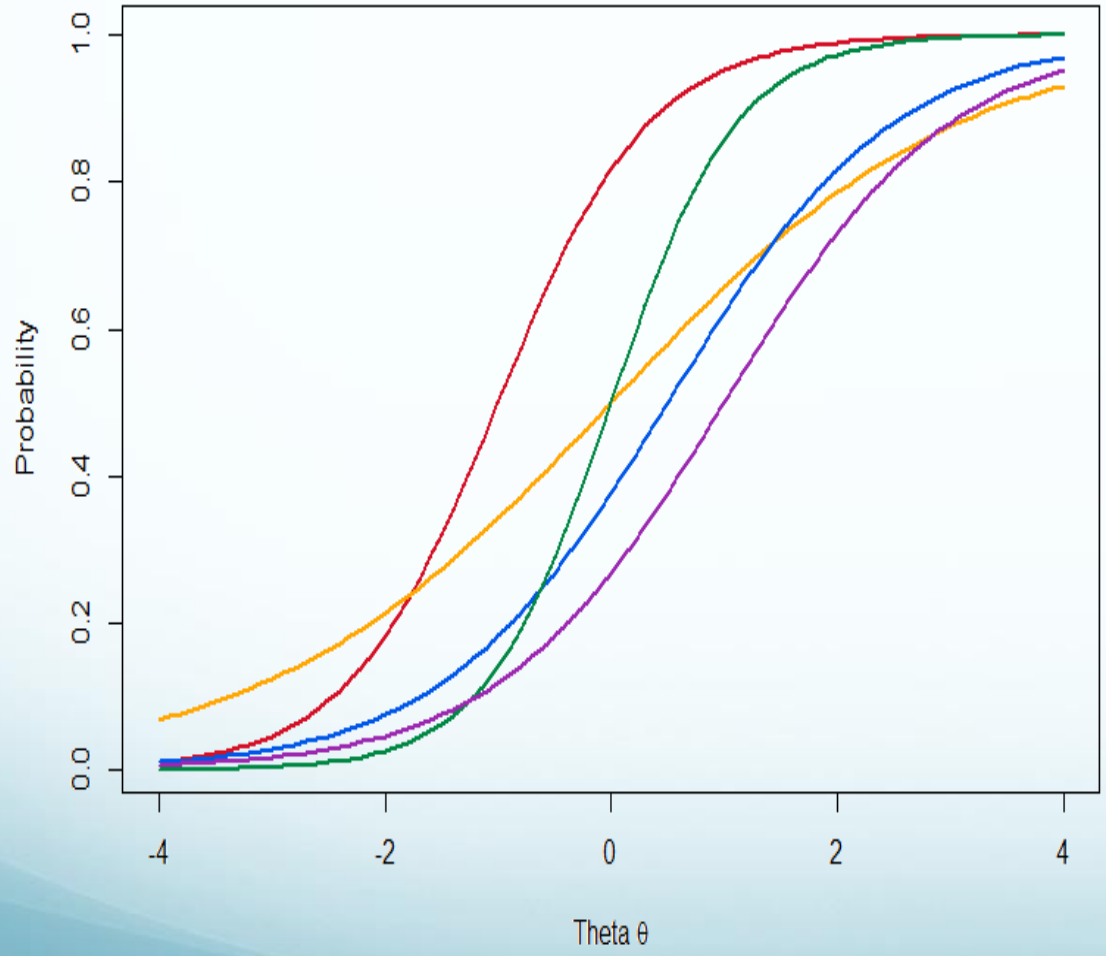


a	b
1	-1
1	-0.05
1	0
1	0.05
1	1

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

CONCERTO

IRT parameters (2pl)

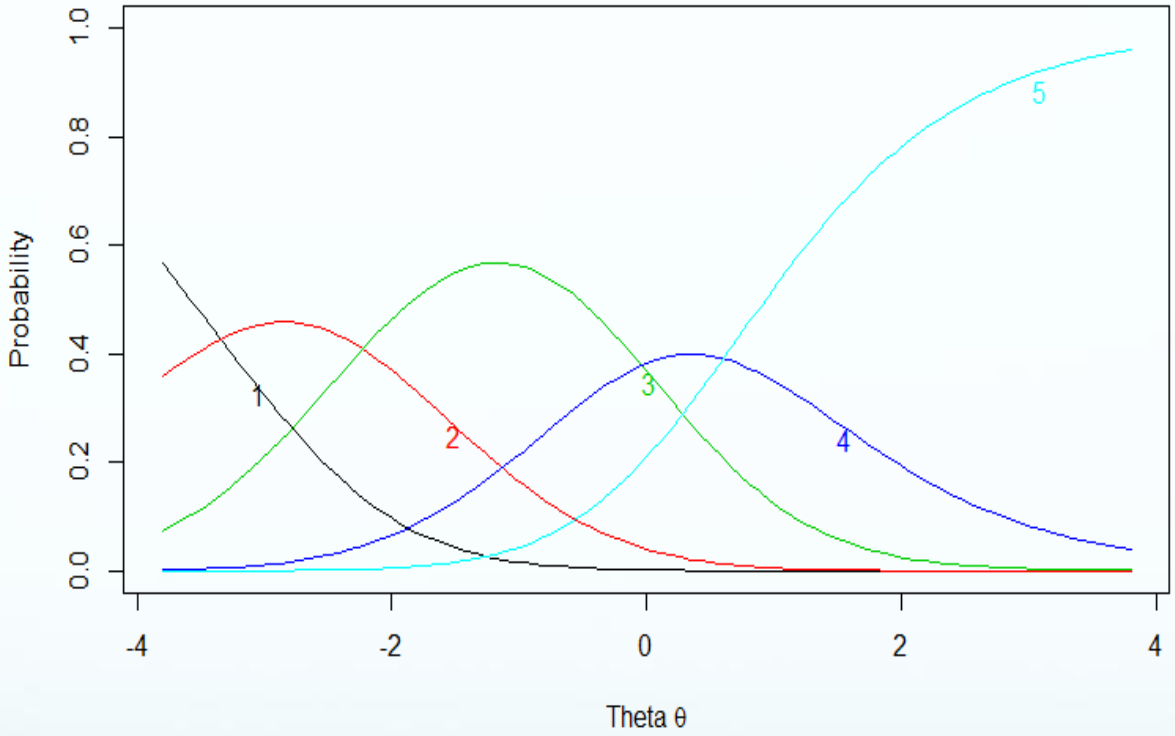


a	b
1	1.25
1.7	-1
0.6	0
2	0
1	1.5

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

CONCERTO

IRT parameters (1pl polytomous)



a	b1	b2	b3	b4	b5
1	-3.75	-3	-1	0.25	1.5

CONCERTO

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-

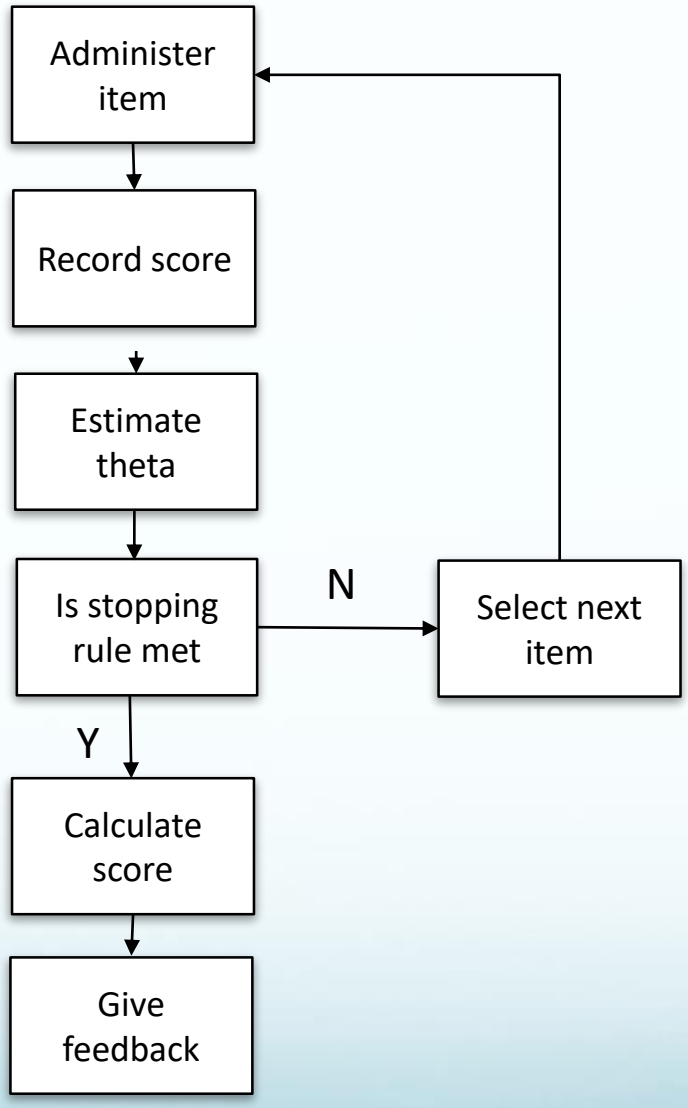
- More flexible
- More efficient
- More accurate
- Better targeted

- 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 assesment
- Integrated feedback


Administer item
Record score
Estimate theta
Is stopping rule met
N
Select next item
Y
Calculate score
Give feedback

Standard questionnaire to assess Lisa

= A question from our questionnaire

?

Maths ability

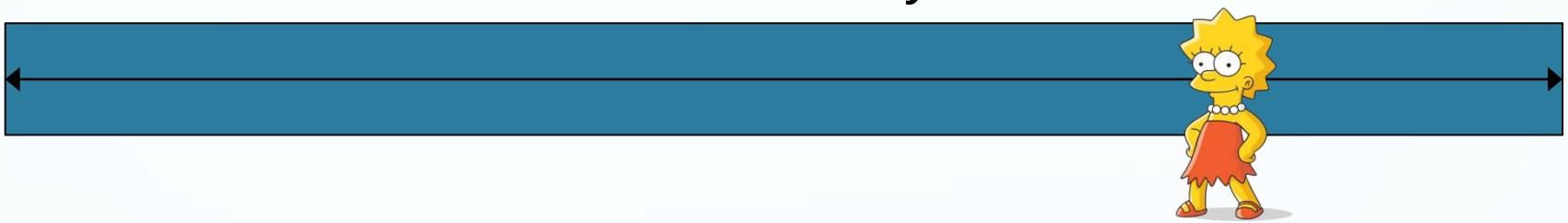
2+2

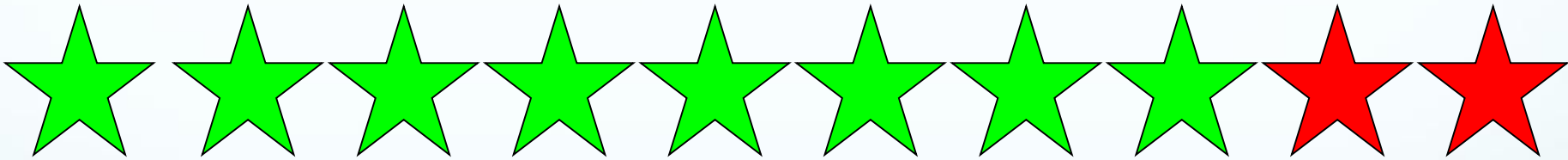
1134 x 16

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

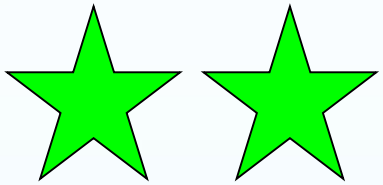
Maths ability

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

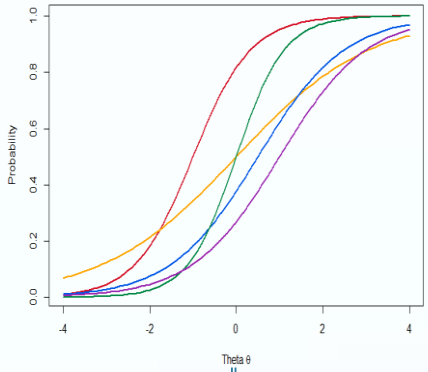
Maths ability

Lisa's ability

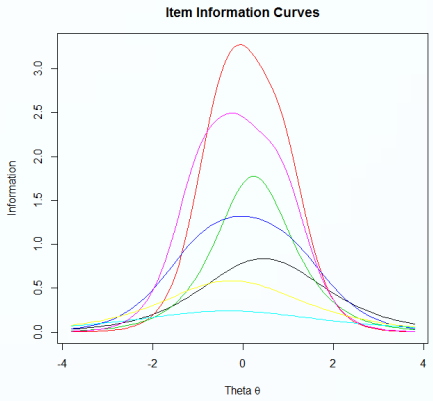
What does a CAT know?

Score input

Item characteristic curves



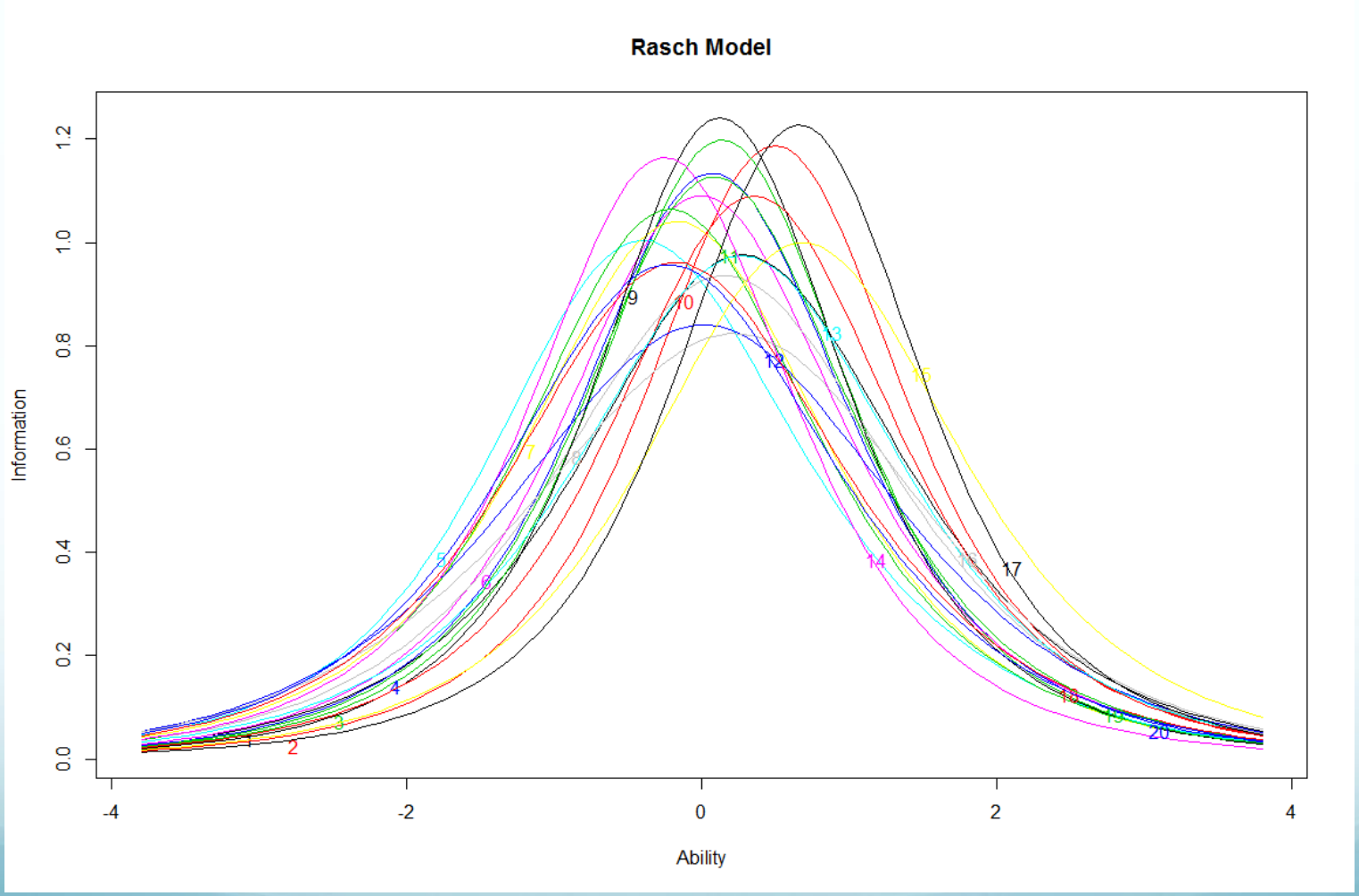
Item information



[Next item, theta estimate, standard error]

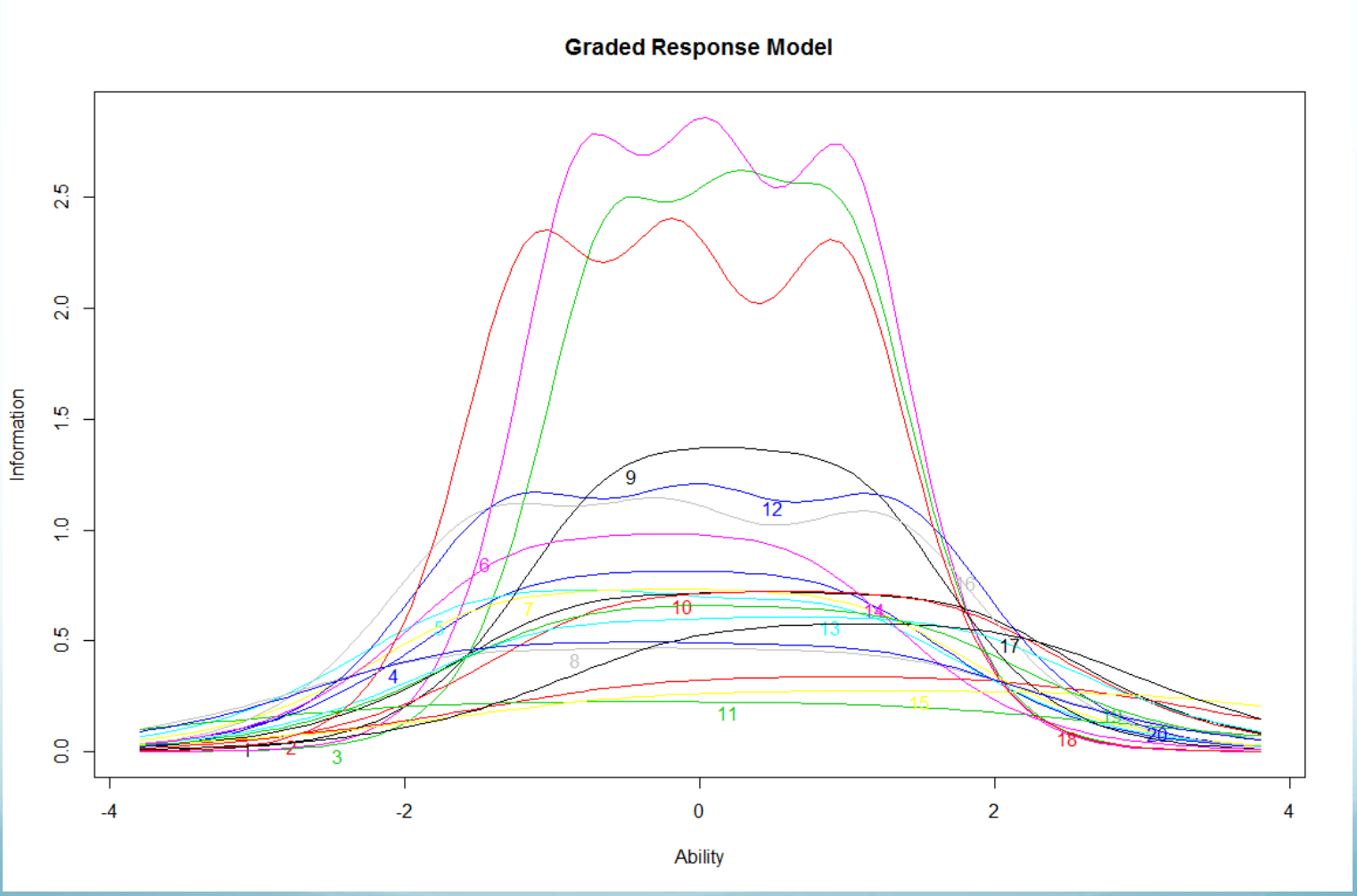
Item Information

- Rasch model



Item Information

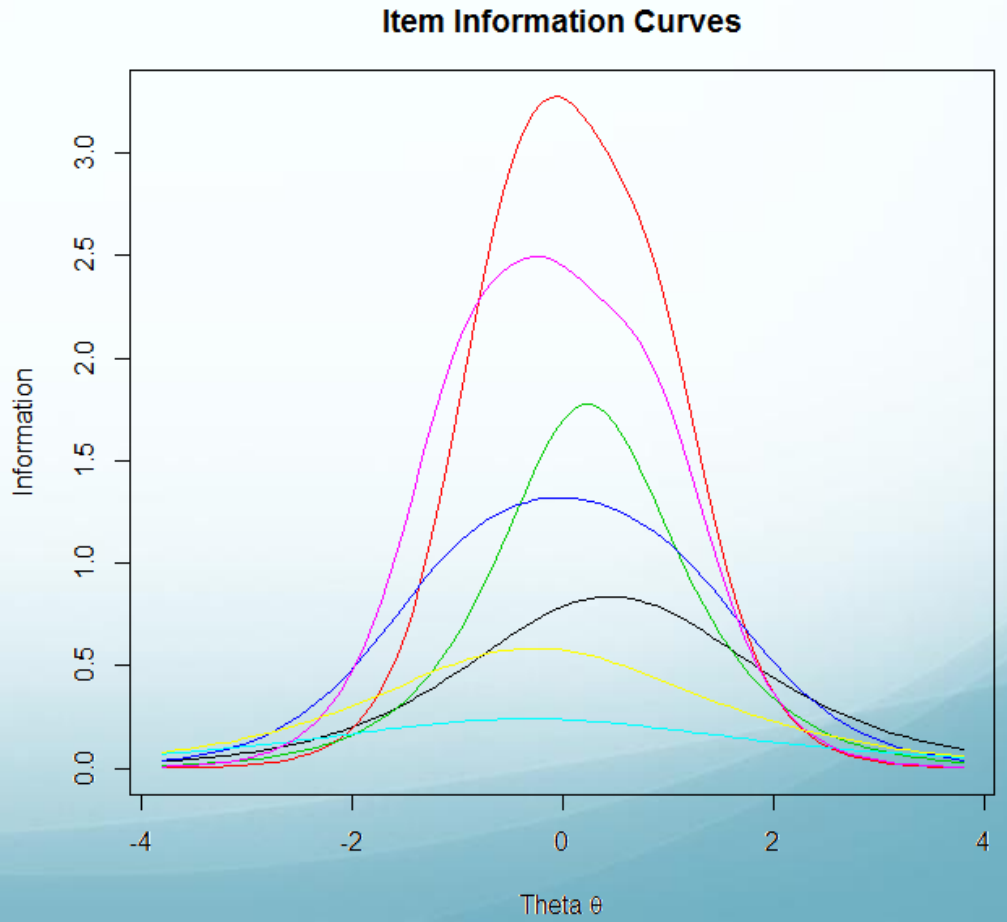
- Graded Response Model



Item Selection

- Maximum information at estimated level of theta

- Bayes Modal (MAP)
- Expected *a posteriori* (EAP)

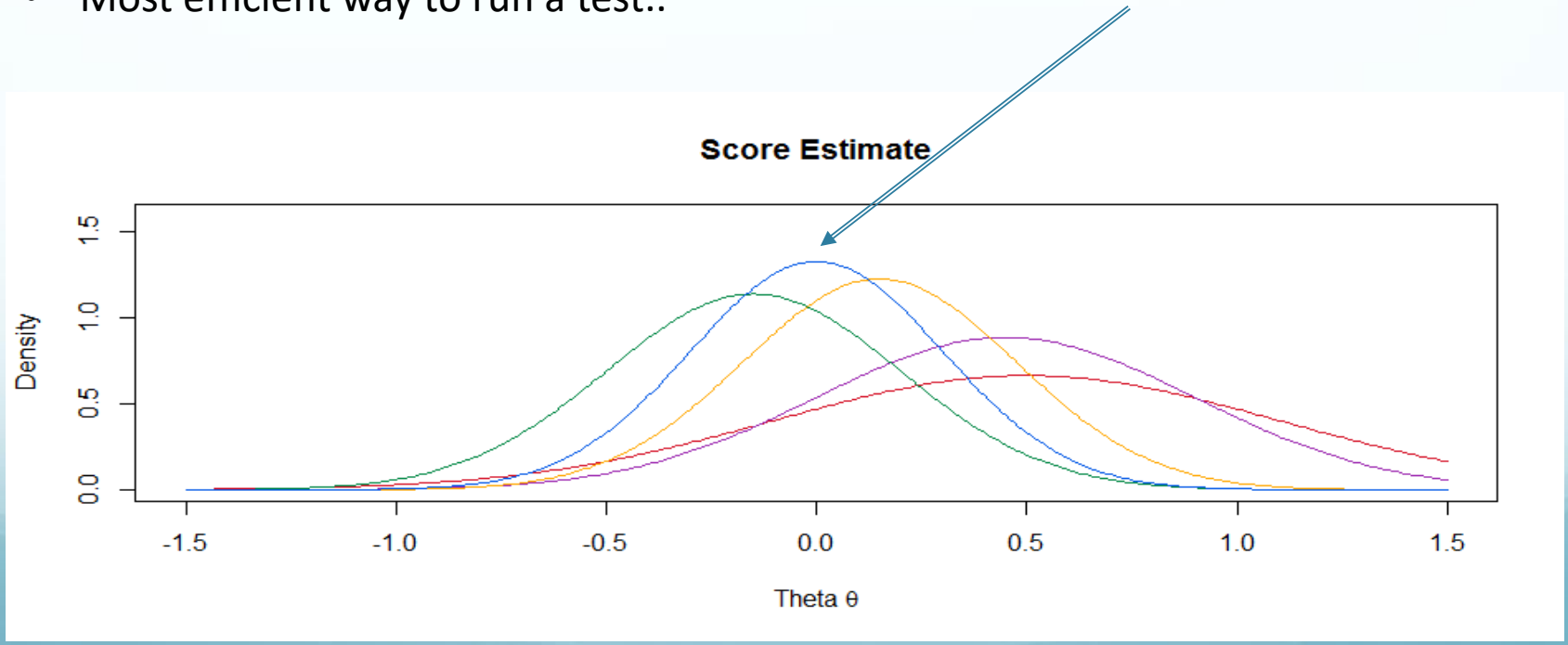


Maximium Information

- Maximum information at estimated level of theta

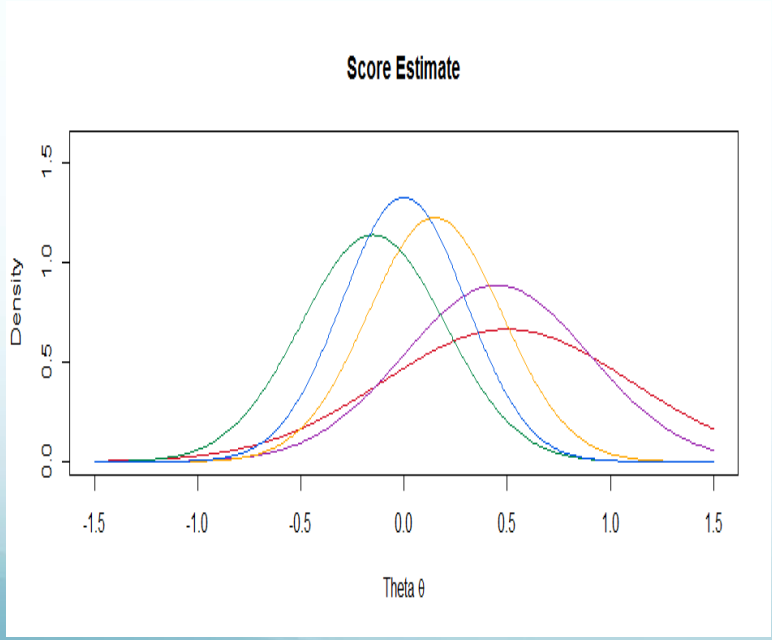
- Proposed by Fisher, developed by Lord
- Each item is selected to provide the maximum information, given the provision estimate of ability (theta)
- Generally items with the steepest discrimination

- Most efficient way to run a test..

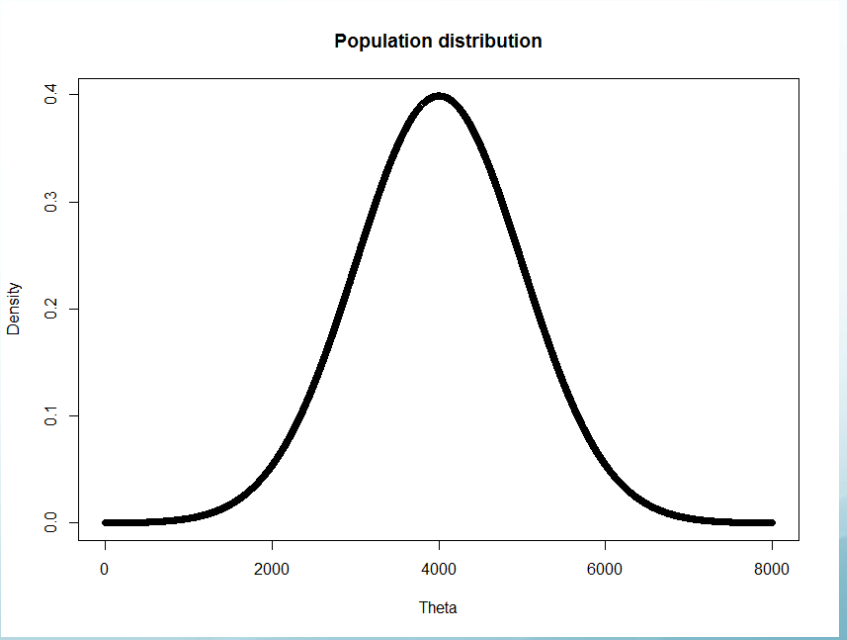


Bayes Modal (MAP)

- 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



X

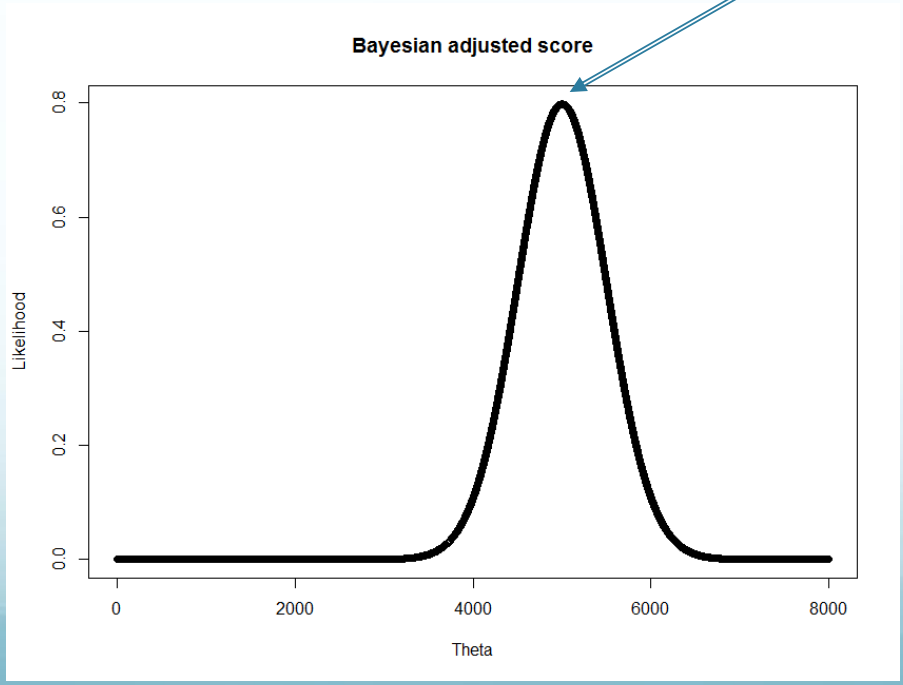


Bayes Modal (MAP)

- 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

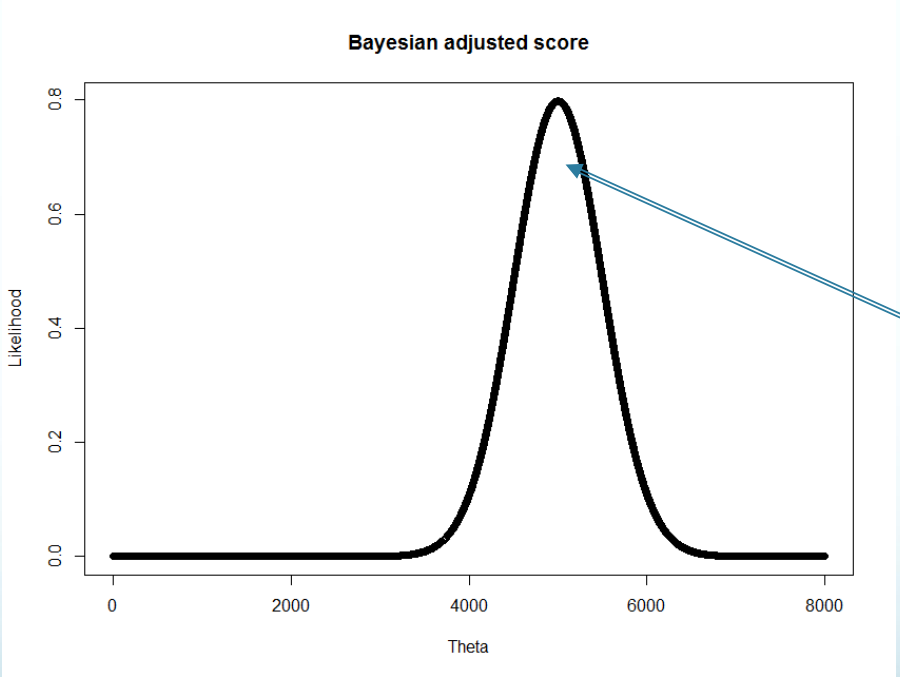
MAP

=



Expected *a posterori* (EAP)

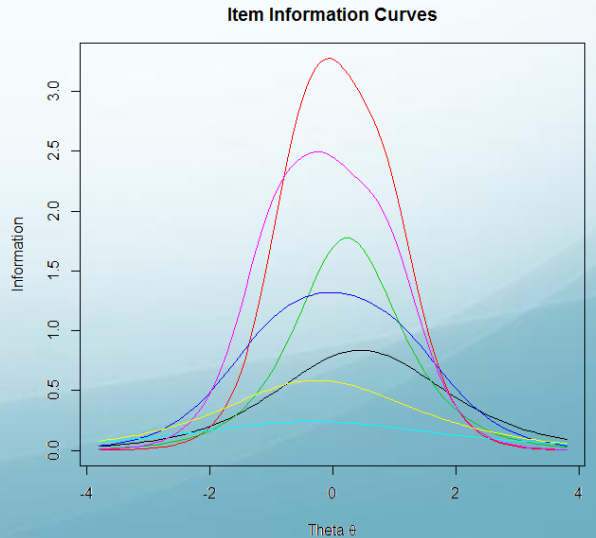
- Instead of taking maximum point of the Bayesian adjusted likelihood function, we take an average value weighted by the EAP function



Average value weighted by function

Specific Information

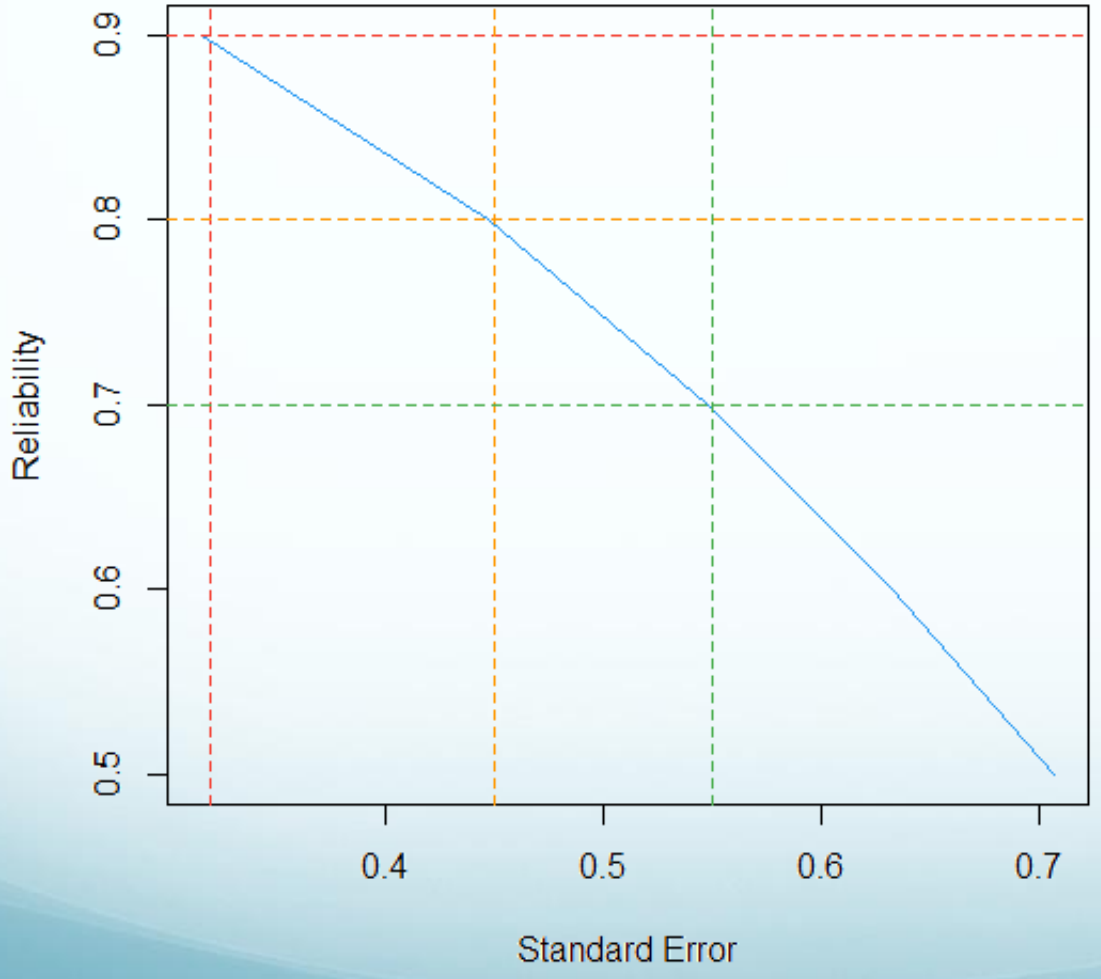
- Proposed by Davey and Fan
- Administer items to achieve pre-selected information targets
- Useful for tests with a cut-off where we don’ t ‘care’ about the information relating to people who are far away from the cut-off



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



Alpha(0.90) = SE(0.32)

Alpha(0.80) = SE(0.45)

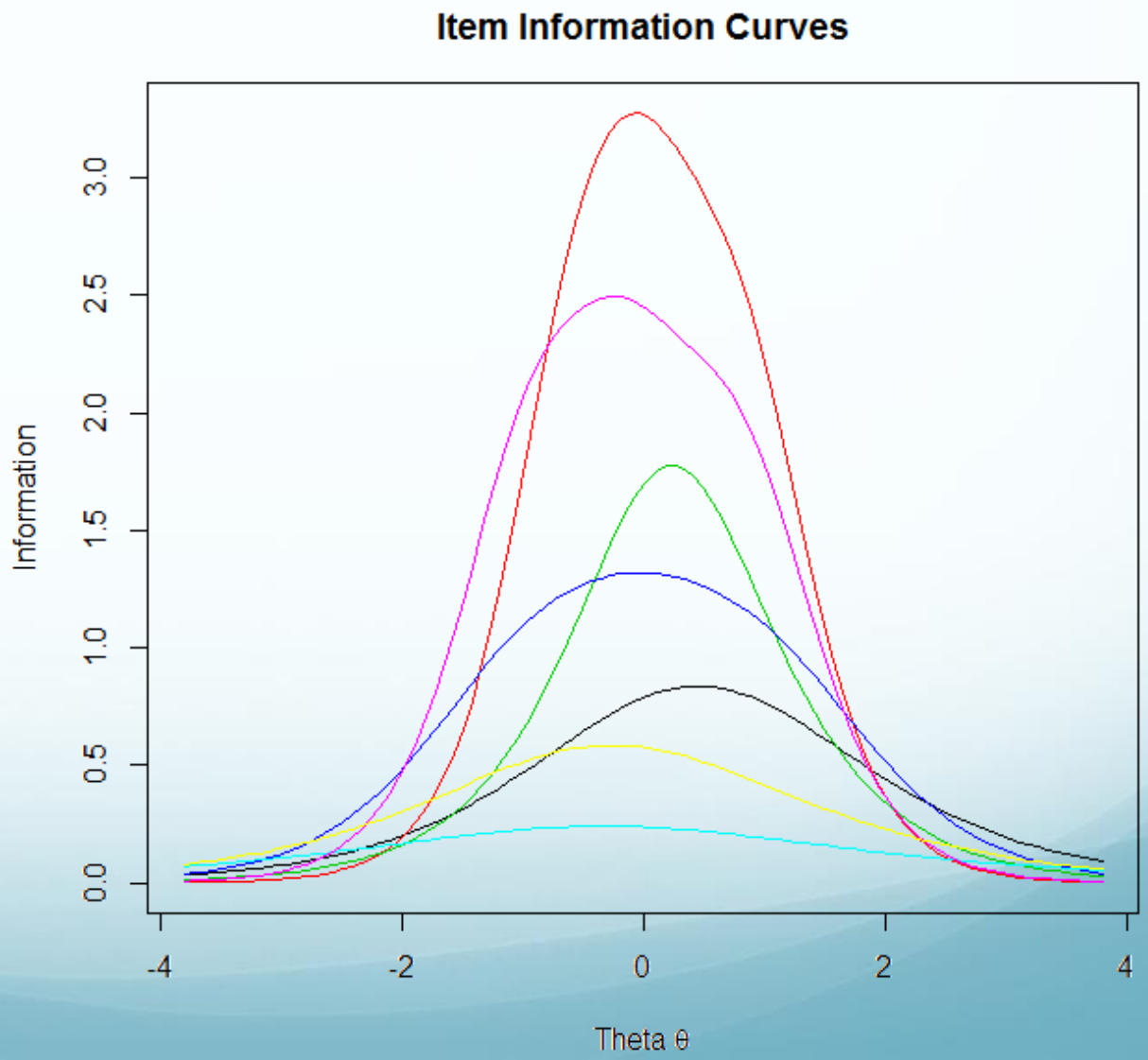
Alpha(0.70) = SE(0.55)

Item Selection quiz..

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

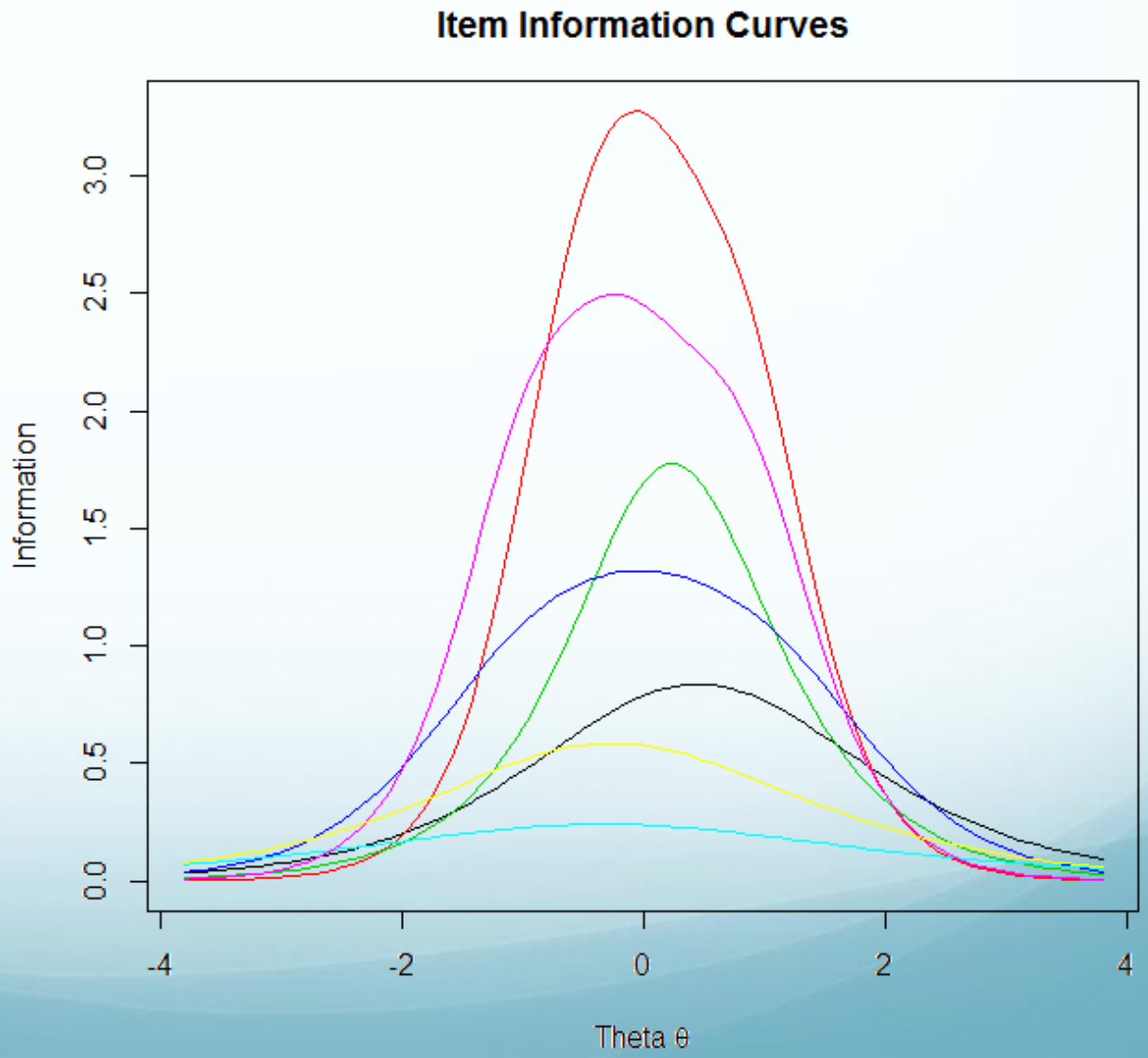
- What additional information would a Bayesian want?

- 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.