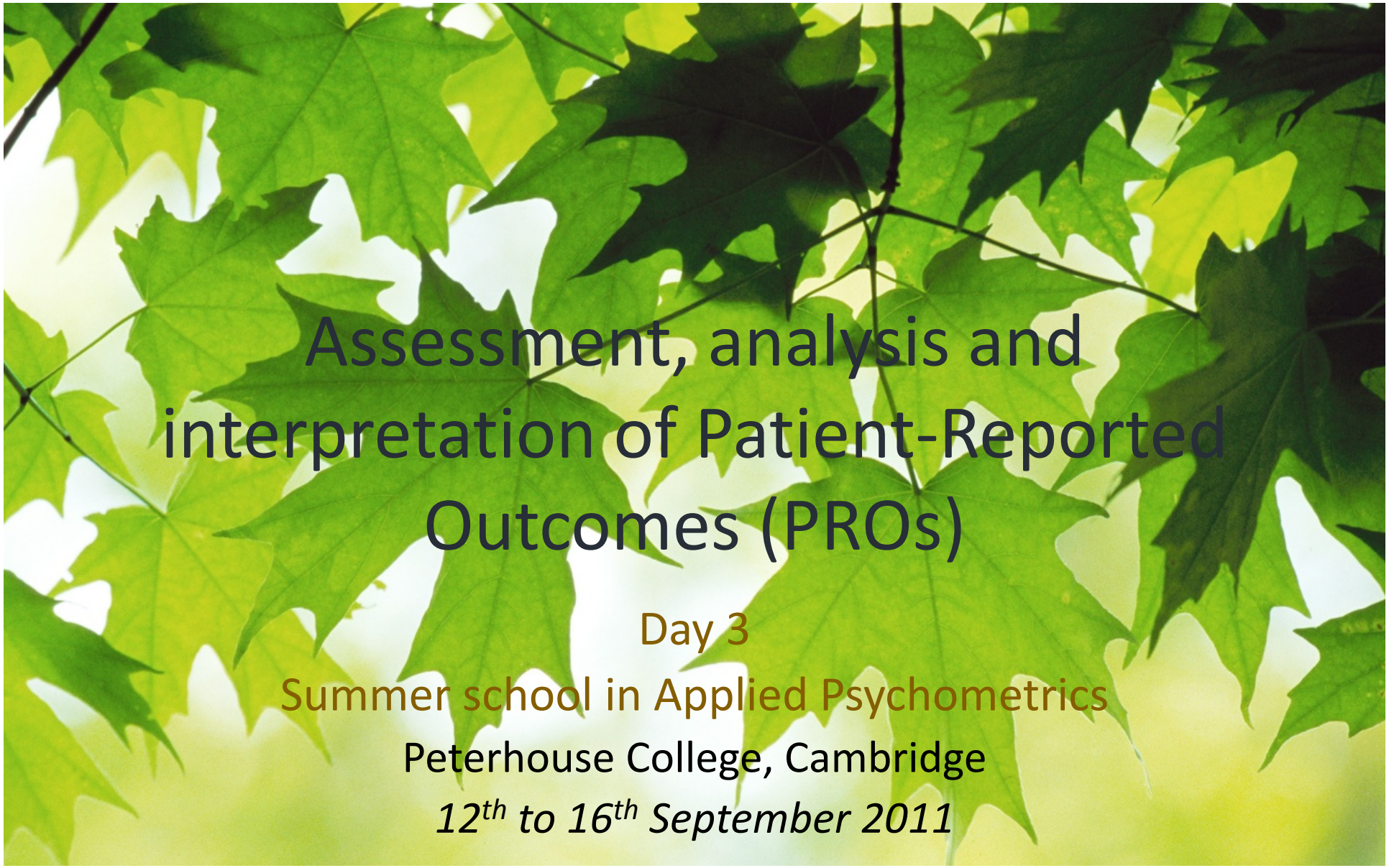


Assessment, analysis and interpretation of Patient-Reported Outcomes (PROs)

Day 3

Summer school in Applied Psychometrics

Peterhouse College, Cambridge

12th to 16th September 2011

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

This course is prepared by

Anna Brown, PhD ab936@medschl.cam.ac.uk
Jan Stochl, PhD js883@cam.ac.uk
Tim Croudace, PhD tjc39@cam.ac.uk
(University of Cambridge, department of Psychiatry)
Jan Boehnke, PhD boehnke@uni-trier.de
(University of Trier, Department of Clinical Psychology and Psychotherapy)

The course is funded by the ESRC RDI and hosted by
The Psychometrics Centre

Jan Boehnke

7. UNIDIMENSIONAL IRT MODELS FOR BINARY DATA

Why unidimensional measures?

"To the degree that one uses a single score from a target measure that includes multiple dimensions (such as a measure of posttraumatic stress disorder thought to include four factors, or a measure of extraversion thought to have six facets), one's construct validation/theory test has theoretical uncertainty built in. Such a test is likely to have reduced scientific value."

Smith, G. T., McCarthy, D. M., & Zapolski, T. C. B. (2009). On the value of homogeneous constructs for construct validation, theory testing, and the description of psychopathology. *Psychological Assessment*, *21*, 272-284.

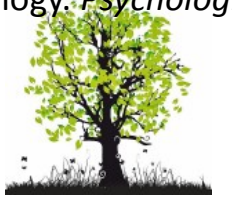
Why are unidimensional measures so interesting?

On the other side, if relationships are found between heterogeneous items / scales and other predictors:

"If such an item predicts a criterion, one will not know which aspect of the item accounts for the covariance. The same reasoning extends to tests: If a test includes multiple dimensions, one cannot know which dimensions account for the test's covariance with measures of other constructs."

Smith, G. T., McCarthy, D. M., & Zapolski, T. C. B. (2009). On the value of homogeneous constructs for construct validation, theory testing, and the description of psychopathology. *Psychological Assessment*, *21*, 272-284.

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre



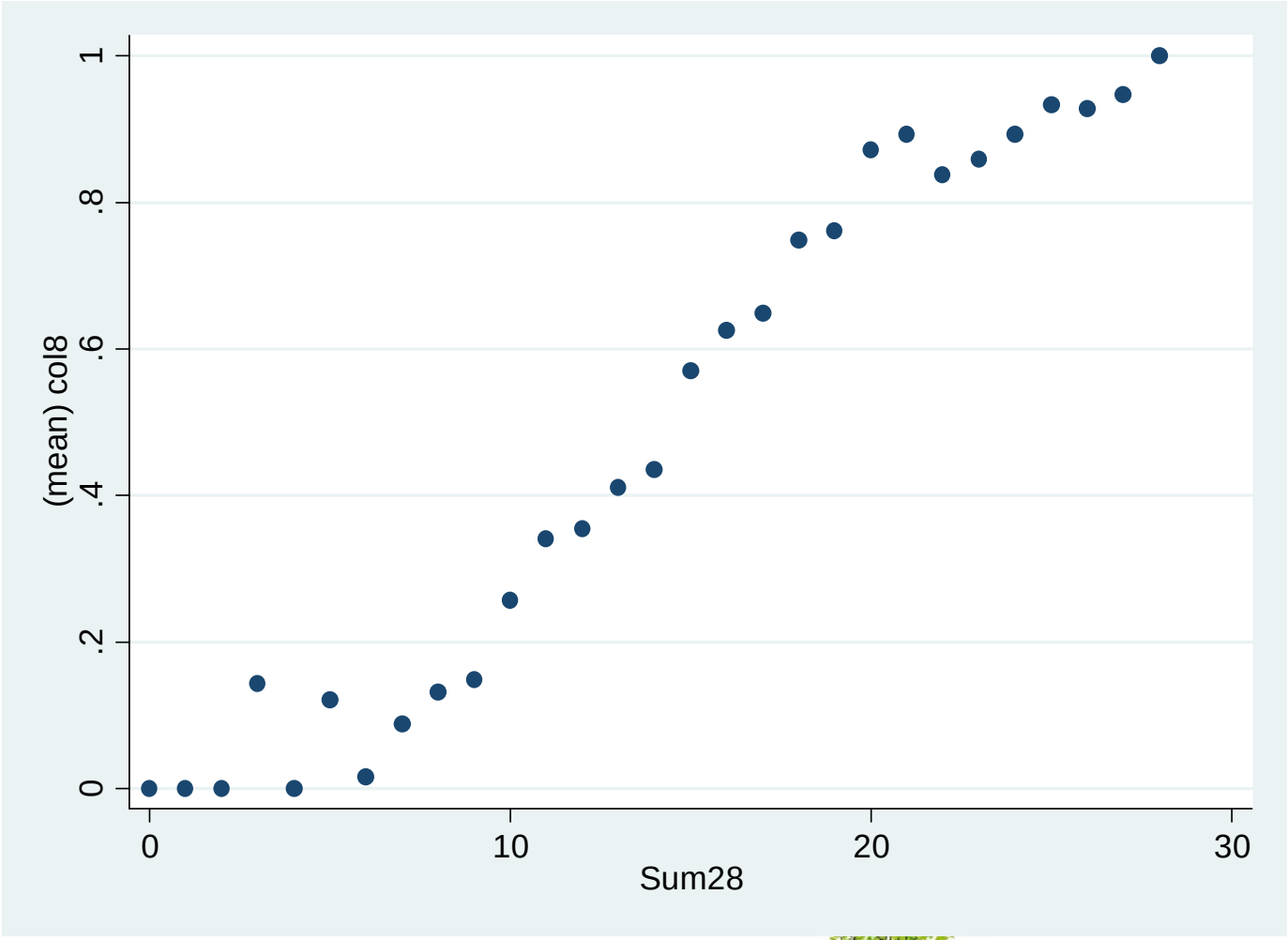
Classical Test Theory

- in CTT it is *assumed* that all items represent the same dimension
- are equally important when representing this dimension
- intervals between the response categories of every item supposed to be the same
- tends to result in items that have the same difficulty

Item Response Theory

- is a generic term for a range of scaling procedures
- all of these model the probability of a response in a specific category of an item ("to solve the item")
- as being dependent on the latent trait ("ability")

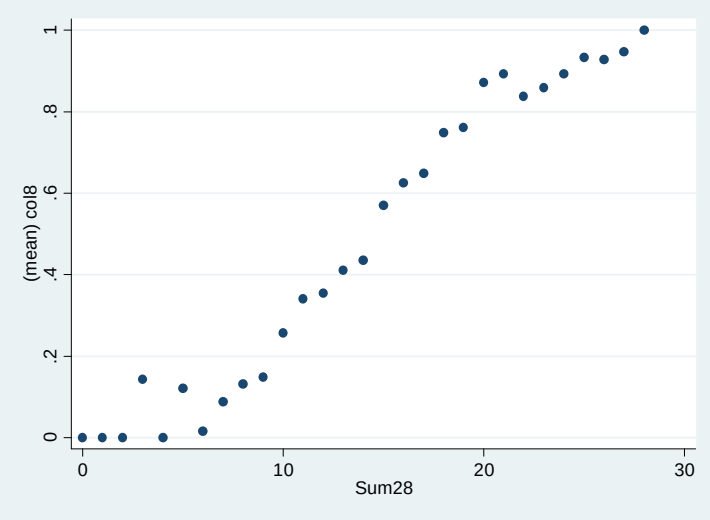
Scores, items, probability



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Scores, items, probability

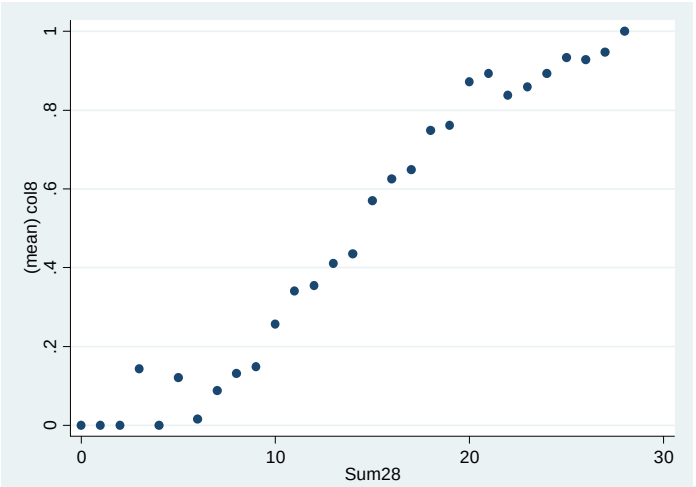
- x-Axis: total score GHQ28
- y-Axis: mean of Item 8 (dich.)

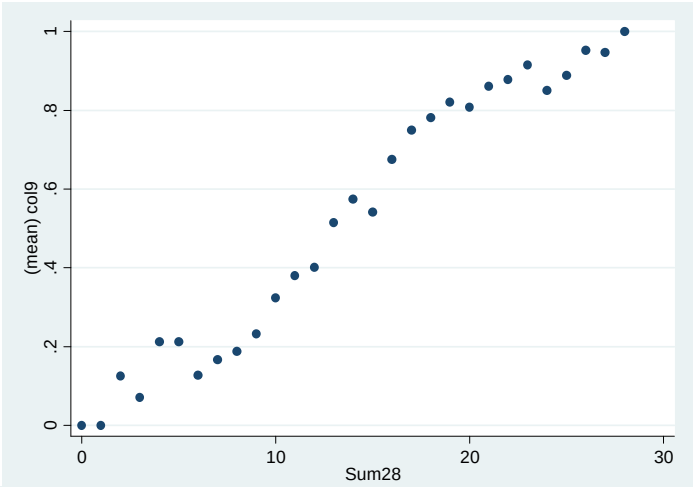


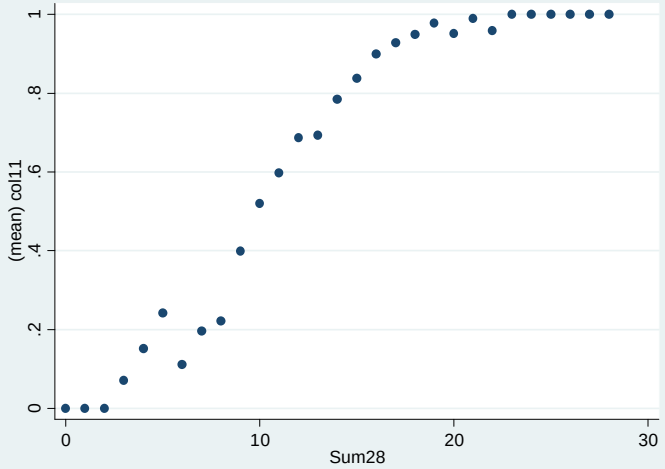
- The picture shows the increase in mean values in item 8 ("lost sleep over worry")
- increase in overall distress hightens the 'probability' of indicating "loss of sleep"

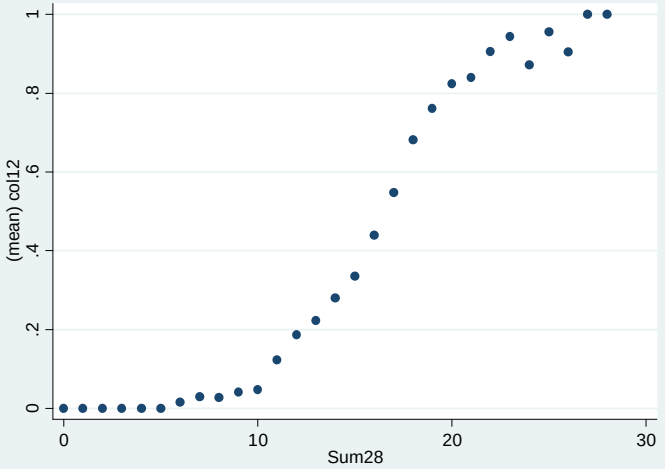
UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Scores, items, probability



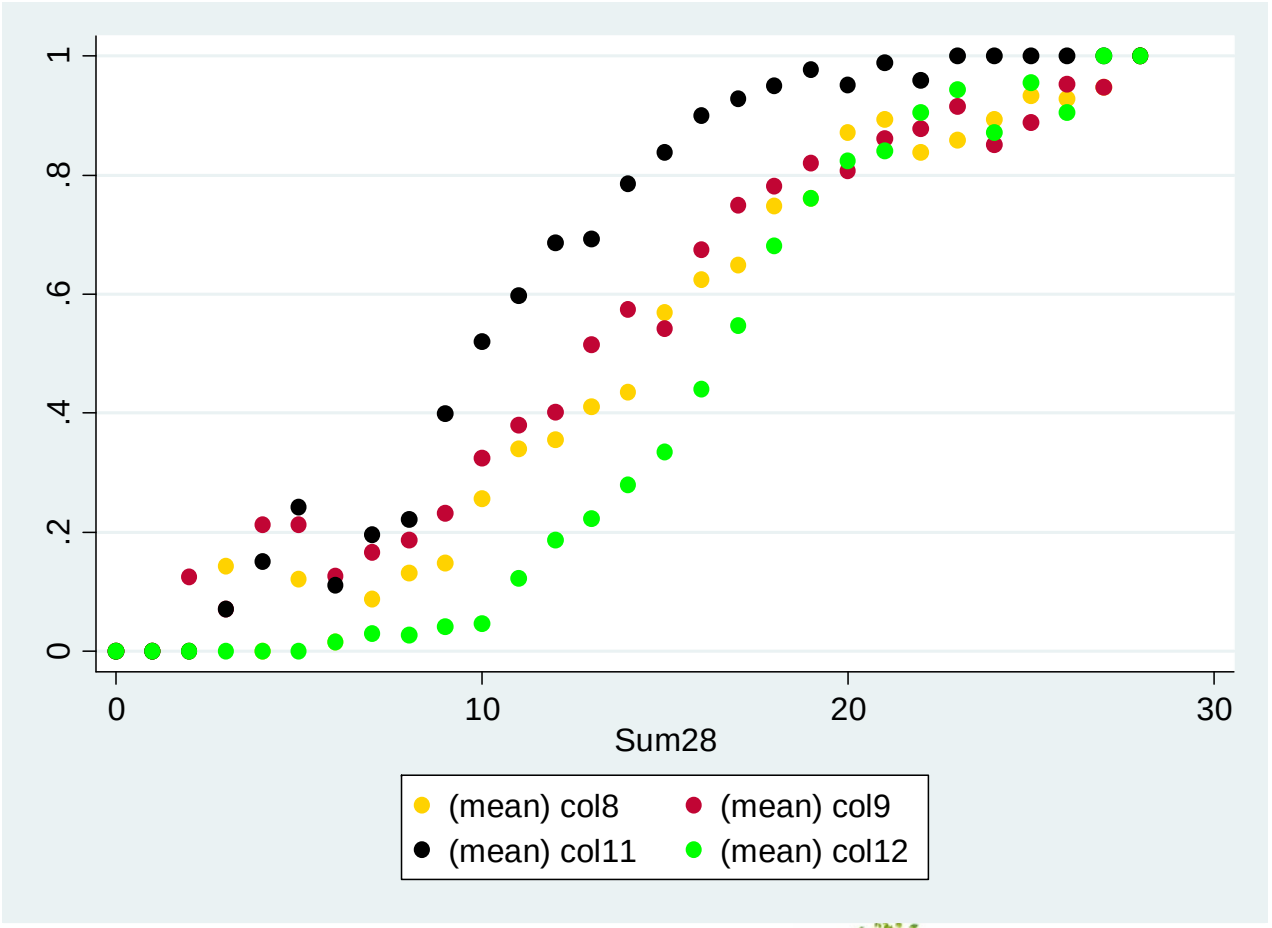




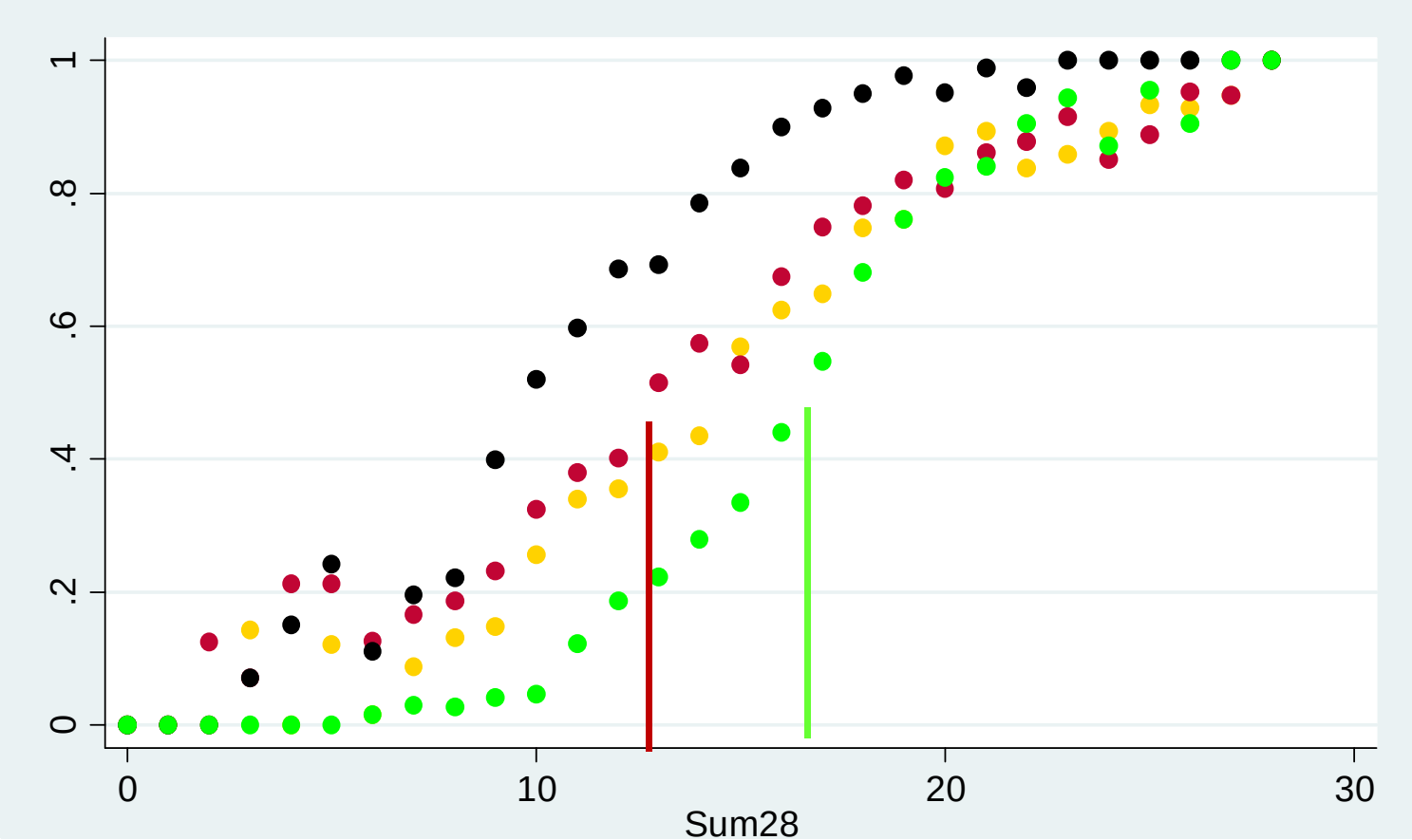


UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Scores, items, probability

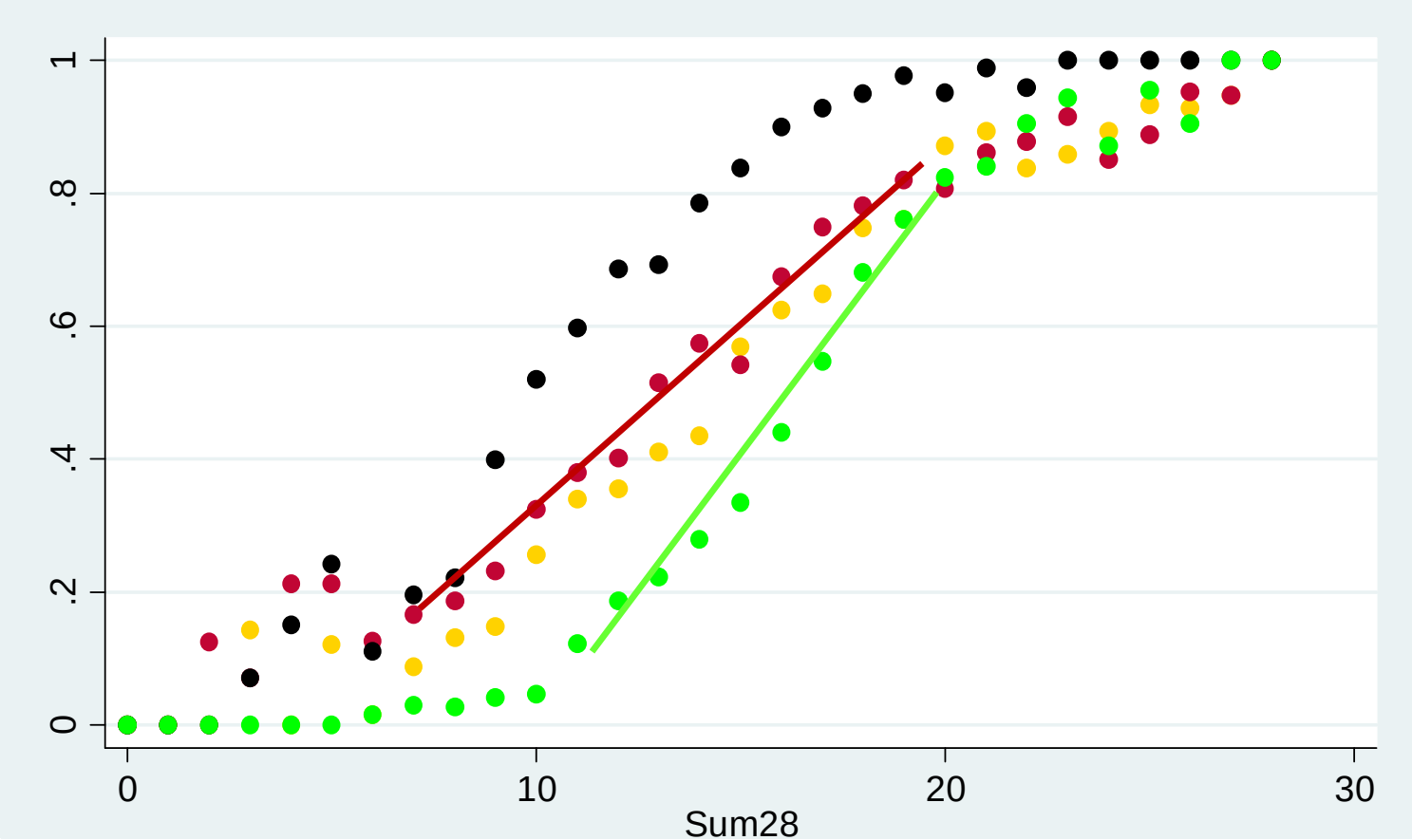


UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre



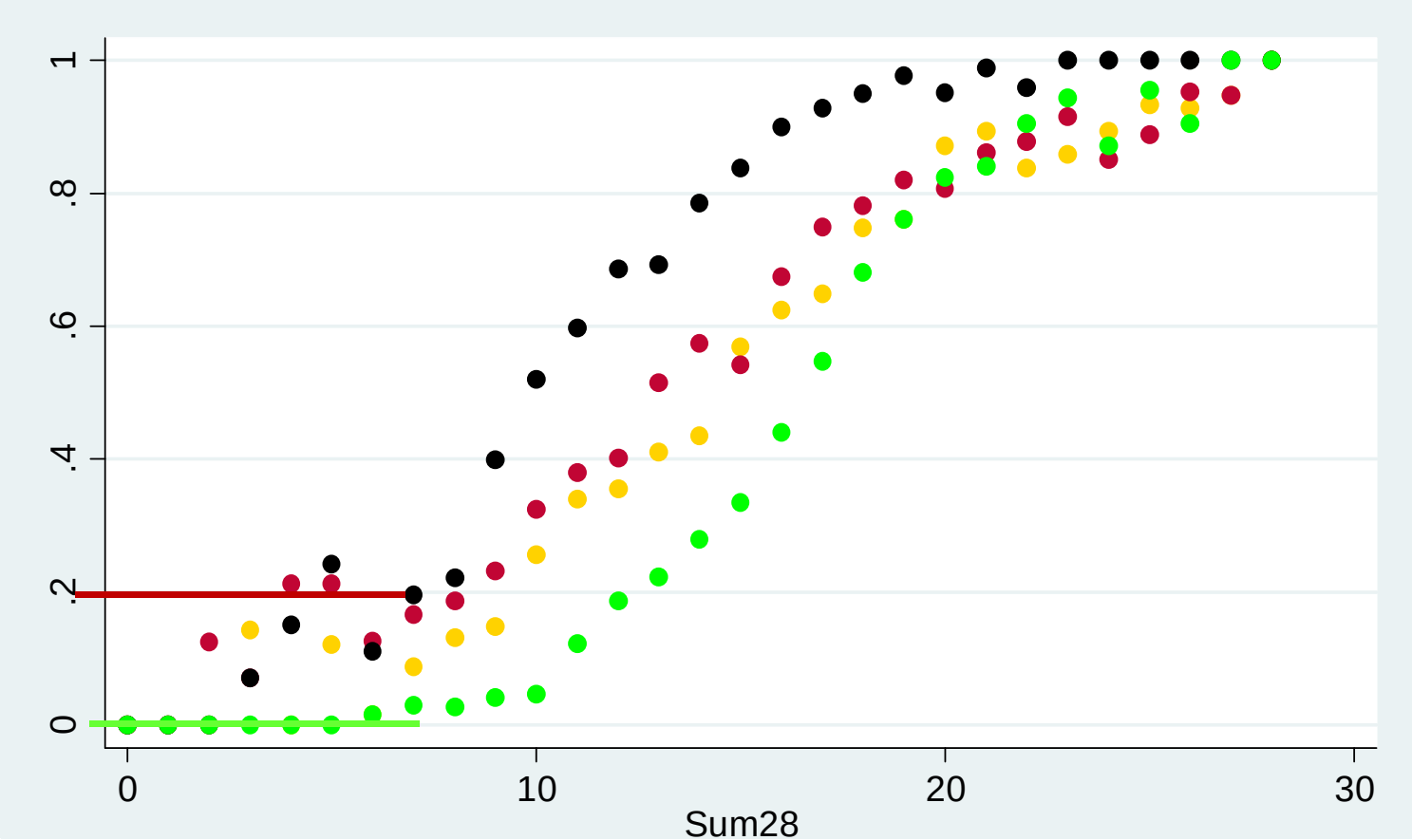
- Items differ in their position on the score, where curvature turns from left to right (will become known as "difficulty")

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre



- Items differ in their slopes (will become known as "discrimination")

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre



- and maybe items differ in their base rate probabilities (will become known as "guessing")

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Item Response Theory

- IRT models now try to estimate instead of the score a latent variable that
  - explains all covariation between the items (unidimensionality / local stochastic independence)
  - and (in a non-literal sense): the item's probability to be solved can be regressed upon (monotonicity)

IRT is a generic term

- most known is a family of the so called "logistic models"
- general form for dichotomous items:

$$P(\theta_h) = c + (1-c)\frac{e^{a(\theta_h - b)}}{1+e^{a(\theta_h - b)}}$$

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

IRT is a generic term

- θ describes the ability of the person (score in the previous example)

$$P(\theta_h) = c + (1-c)\frac{e^{a(\theta_h - b)}}{1 + e^{a(\theta_h - b)}}$$

- the other three describe the other possible variations:
  - b is the difficulty / position of the item on the latent trait
  - a is the discrimination / slope of the curve
  - c models the base rate / guessing

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Polytomous IRT

- Fayers and Machin write in their book that "ploytomous IRT" not fully developed
- this was already not true in 2007, when the book was published (or 2006, when the final draft came about)
- we will start with dichotomies (simple) but move to polytomous items later

Assumptions of IRT models

- The assumptions of the IRT models covered today are
  - *monotonically* increasing probabilities for the items to be solved depending on the latent trait
  - *unidimensionality*: there is only one trait explaining the associations between the items
  - *local independence*: when the latent trait is controlled for, there are no associations between the items left

And what now?

PRACTICALS:
IRT MODELING USING R

Package "ltm"

- The package "ltm" (Rizopoulos, 2006) was buildt
  - to estimate a range of IRT models
  - for dichotomous as well as polytomous items
  - to provide graphics
  - and tests to check model fit
- to load "ltm" type `library(ltm)`

- The package provides three functions for dichotomous data:
  - `rasch()` for 1PL
  - `ltm()` for 2PL
  - `tpm()` for 3PL
- and two for polytomous data:
  - `gpcm()`
  - `grm()`

Rizopoulos, D. (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses, Journal of Statistical Software, 17 (5), 1-25.

Reading data

- R is a console, so no "GUI" / point-&-click surface is directly available

- therefore: everything in code, e.g.: reading data:

```
GHQ28 <- read.table(file.choose(),
  header=TRUE, sep="\t",
  na.strings="NA", dec=".",
  strip.white=TRUE)
```

Estimating models: the Anxiety subscale of the GHQ

- two options to choose the Anxiety/Insomnia scale items:
  - adress only those columns of the data file and write them into a new object:

```
Anxiety<-GHQ28[ ,8:14]
```

  - attach the data file:

```
attach(GHQ28)
Anxiety<-cbind(anxi1,
  anxi2,anxi3,anxi4,anxi5,anxi6,anxi7)
```

To estimate a 1PL model

- this model only uses one parameter in addition to the ability:
  - the items differ in their difficulty

$$P(\theta_h) = \frac{e^{(\theta_h - b)}}{1 + e^{(\theta_h - b)}}$$

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

To estimate a 1PL model

- Package description:

```
rasch(data, constraint = NULL, IRT.param = TRUE, start.val = NULL,
    na.action = NULL, control = list(), Hessian = TRUE)
```

- "rasch()" is the function that will estimate a number of 1PL models
- the above provides the structure of the command with all options

To estimate a 1PL model

- Looking in the "Example" section provides us with:

```
## The Rasch model for the LSAT data with
## unconstraint discrimination parameter
rasch(LSAT)
```

- for an unconstrined 1PL model only the data matrix has to be specified on which it should be calculated
- in our case instead of "LSAT" just "Anxiety"

To estimate a 1PL model

```
Result1PL<-rasch(Anxiety)
```

- the results of the model estimation will hereby saved into a new object, "Result1PL"
- to look at them:
  - type `Result1PL` in the console
  - or `summary(Result1PL)`

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

To estimate a 1PL model

```
summary(Result1PL)

Model Summary:
   log.Lik       AIC       BIC
 -11002.49 22020.99 22068.77

Coefficients:
               value std.err   z.vals
Dffclt.anxi1  0.0023  0.0303   0.0774
Dffclt.anxi2 -0.1690  0.0305  -5.5458
Dffclt.anxi3 -0.8603  0.0351 -24.5105
Dffclt.anxi4 -0.7408  0.0339 -21.8767
Dffclt.anxi5  0.3782  0.0314  12.0599
Dffclt.anxi6 -0.1261  0.0304  -4.1468
Dffclt.anxi7  0.3475  0.0312  11.1364
Dscrmn        2.0544  0.0465  44.2081
```

Difficulty = item's position on the latent trait

Discrimination = how strongly do the items discriminate between high and low trait values

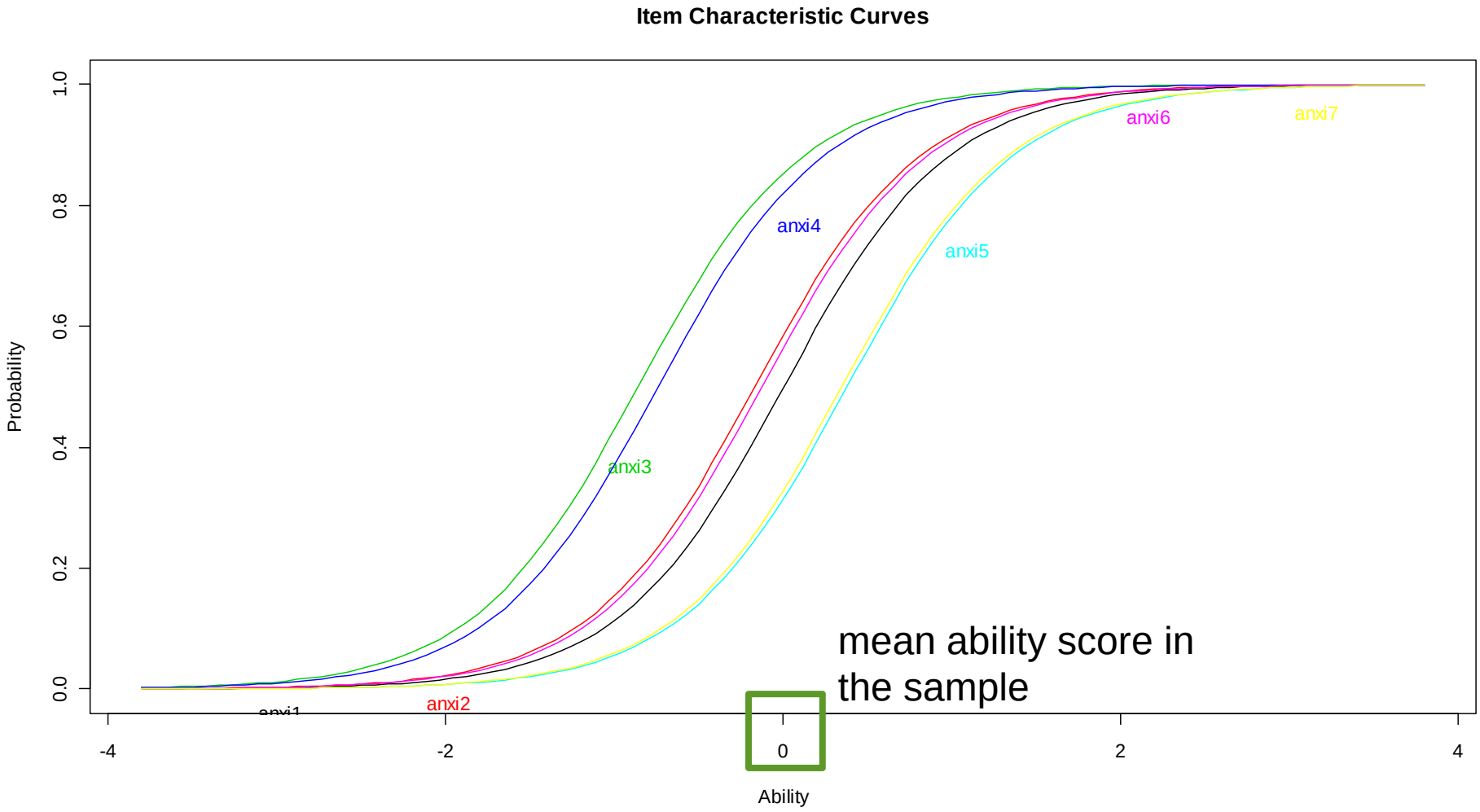
UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Looking at estimates from 1PL

- results from the estimation can directly plotted by using `plot(Result1PL)`

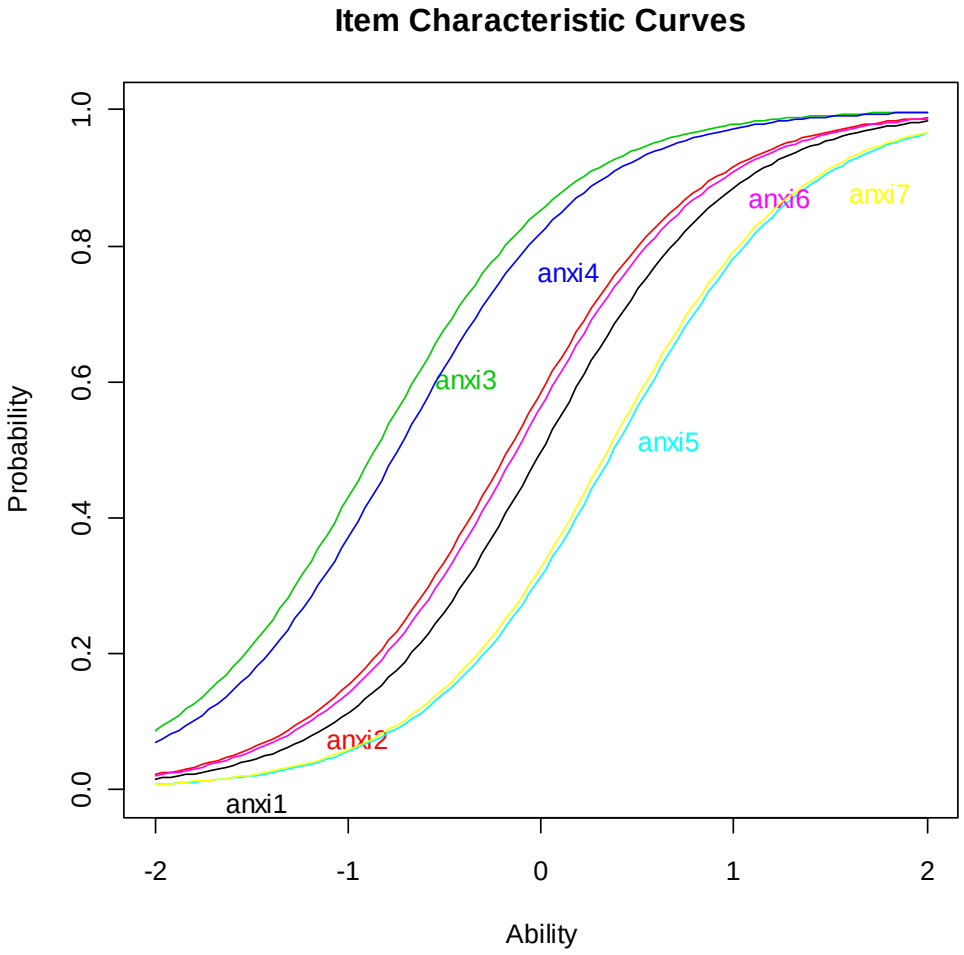
```
plot(x, type = c("ICC", "IIC"), items = NULL,
     zrange = c(-3.8, 3.8), z = seq(zrange[1], zrange[2], length = 100),
     annot, labels = NULL, legend = FALSE, cx = "topleft", cy = NULL,
     ncol = 1, bty = "n", col = palette(), lty = 1, pch, xlab, ylab,
     main, sub = NULL, cex = par("cex"), cex.lab = par("cex.lab"),
     cex.main = par("cex.main"), cex.sub = par("cex.sub"),
     cex.axis = par("cex.axis"), plot = TRUE, ...)
```

Looking at estimates from 1PL



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

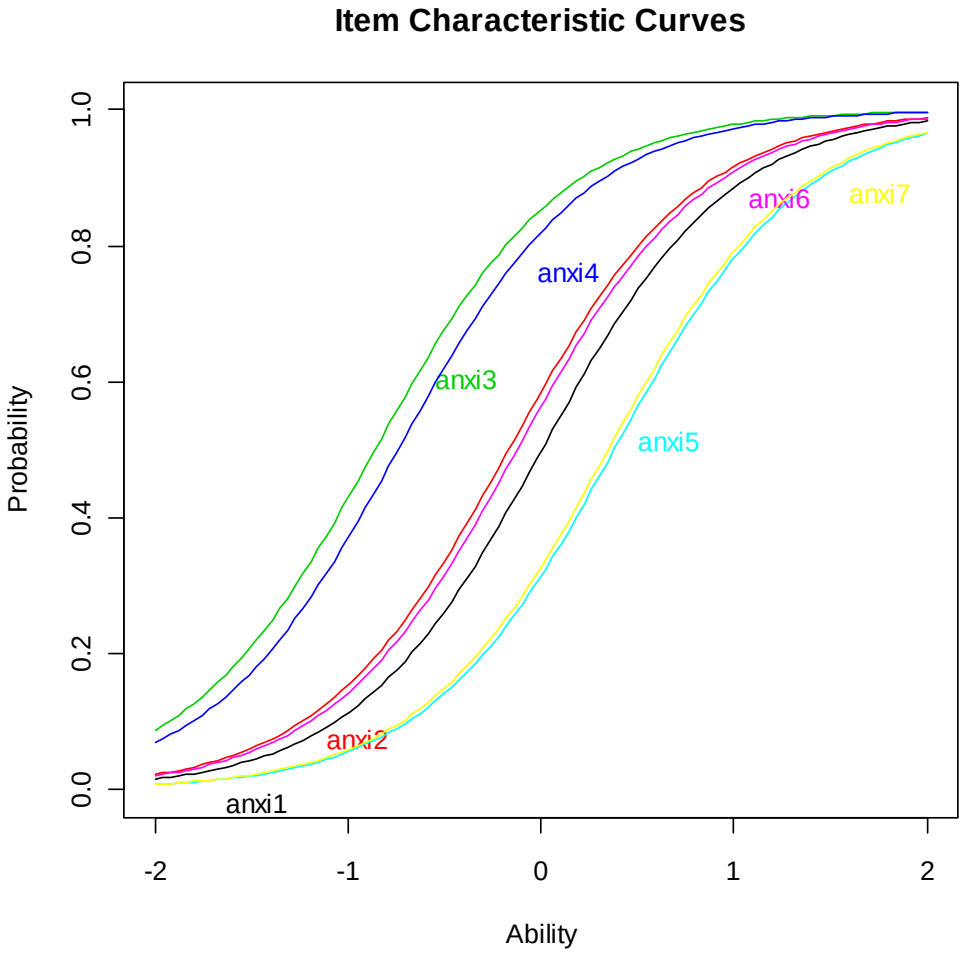
Looking at estimates from 1PL



- items differ in their difficulty:
  - easiest (b = -.86): anxi3 / GHQ16 "constant strain"
  - most difficult (b = .38): anxi5 / GHQ19 "panicky / scared"

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Looking at estimates from 1PL



- 1PL only estimates b for every item, therefore the other parameters are:
  - same slope with a = 2.05 for all items
  - no guessing parameter, for all items c = 0

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Testing 1PL

- Question: does the model describe the data sufficiently?
- For Rasch models a simulation / bootstrap test is at hand: `GoF.rasch()`

Testing 1PL

- Test procedure:
  - the test compares observed and predicted numbers of response patterns (Pearson-χ^2):

$$\sum_{r=1}^{2^p} \frac{\{O(r) - E(r)\}^2}{E(r)},$$

- unfortunately, this test only follows in (very,...) large samples a χ^2-distribution
- therefore a parametric bootstrap is applied

Testing 1PL

- Bootstrap:
  - the parametric bootstrap simulates *B* samples based on the estimated parameters
  - in these simulated samples the model is estimated each time and the respective χ^2 value computed and saved
  - from these *B* samples the actual significance is calculated by $[1 + \sum_{i=1}^{B} I(T_i > T_{obs})]/(B+1)$

Testing 1PL

- Code in our specific case:

```
Test1PL<-GoF.rasch(Result1PL,B=100)
Test1PL
```

  - observed χ^2 = Tobs = 495.21
  - bootstrapped p-value: 0.01

- 1PL seems not to be enough to explain the responses

Testing 1PL: dimensionality

- One of the assumptions of IRT models is "unidimensionality"
- after controlling for the latent dimension therefore no correlations between the items should exist
- test with parametric bootstrap / parallel analysis: is the second eigenvalue of a factor analysis of the tetrachoric correlations higher than expected?

Testing 1PL: dimensionality

- After the estimation model parameters (here: difficulties) are saved
- with these B samples of the same size are simulated and in each of them a factor analysis of the tetrachoric correlation matrix is done
- the second eigenvalue is saved for each of these runs and it is then tested whether the observed second eigenvalue is higher than expected by the model: $\left(1+\sum_{b=1}^{B} I(T_b \geq T_{obs})\right)/(1+B)$

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Testing 1PL: dimensionality

- Result in our case:

```
Dimen1PL<-unidimTest(Result1PL,Anxiety,B=499)
```

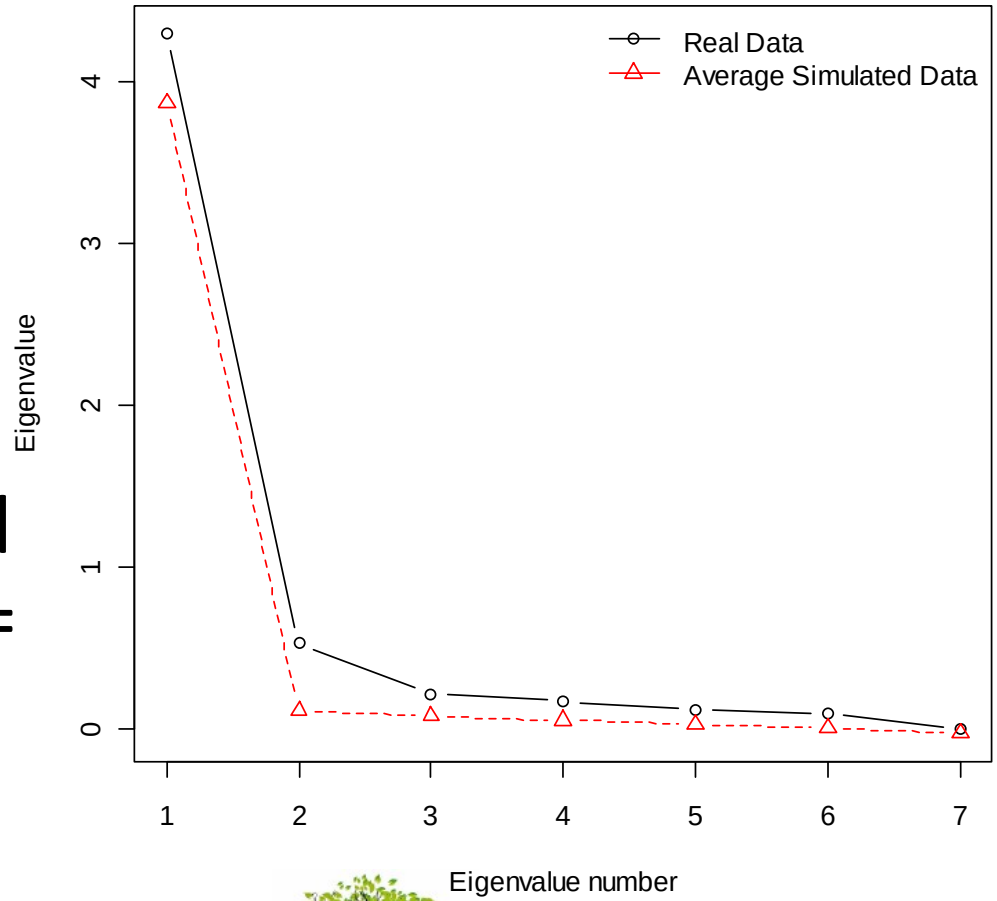
- Alternative hypothesis: the second eigenvalue of the observed data is substantially larger than the second eigenvalue of data under the assumed IRT model

Second eigenvalue in the observed data: 0.5279
Average of second eigenvalues in Monte Carlo samples: 0.1134
Monte Carlo samples: 499
p-value: 0.002

Testing 1PL: dimensionality

```
plot(Dimen1PL,
   type="b",pch=1:2)

legend("topright",
   c("Real Data",
   "Average Simulated
   Data"), lty = 1,pch =
   1:2, col = 1:2, bty =
   "n")
```



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Adding complexity I: The 2PL model

DO ITEMS VARY IN THEIR DISCRIMINATORY POWER?

To estimate a 2PL model

- this model only uses two parameters in addition to the ability:
  - the items differ in their difficulty
  - and also their slopes / discrimination (a)

$$P(\theta_h) = \frac{e^{a(\theta_h - b)}}{1 + e^{a(\theta_h - b)}}$$

To estimate a 2PL model

- Go back to package description:

```
ltm(formula, constraint = NULL, IRT.param, start.val,
    na.action = NULL, control = list())
```

- "ltm()" is the function that will estimate a number of 2PL models
- the above provides the structure of the command with all options

To estimate a 2PL model

- for an unconstrained 2PL model only the data matrix has to be specified on which it should be calculated as well as the number of dimensions to be extracted:

```
Result2PL<-ltm(Anxiety ~ z1)
```

To estimate a 2PL model

```
summary(Result2PL)

log.Lik      AIC      BIC
 -10919.61 21867.22 21950.84

Coefficients:
                value std.err   z.vals
Dffclt.anxi1   0.0076  0.0334   0.2283
Dffclt.anxi2  -0.1956  0.0367  -5.3242
Dffclt.anxi3  -0.8396  0.0378 -22.2130
Dffclt.anxi4  -0.7521  0.0378 -19.8858
Dffclt.anxi5   0.3862  0.0325  11.8967
Dffclt.anxi6  -0.1084  0.0267  -4.0628
Dffclt.anxi7   0.3078  0.0277  11.1136
Dscrmn.anxi1   1.6496  0.0840  19.6454
Dscrmn.anxi2   1.4152  0.0740  19.1211
Dscrmn.anxi3   2.1691  0.1230  17.6346
Dscrmn.anxi4   1.9571  0.1066  18.3600
Dscrmn.anxi5   1.9880  0.1027  19.3576
Dscrmn.anxi6   3.0993  0.1855  16.7093
Dscrmn.anxi7   3.0537  0.1811  16.8654
```

Difficulty = item's position on the latent trait

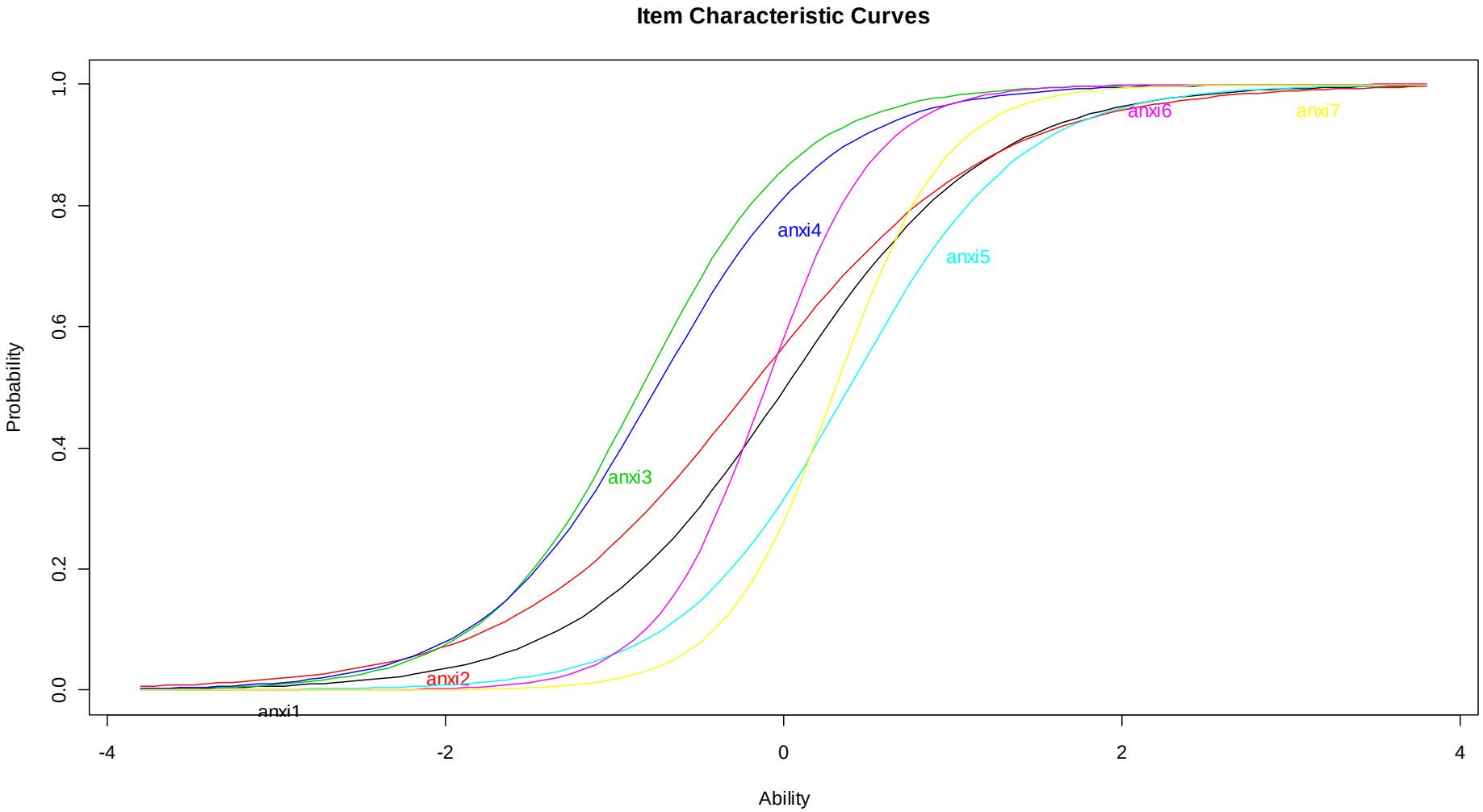
Discrimination = how strongly do the items discriminate between high and low trait values

Looking at estimates from 2PL

- results from the estimation can also directly plotted by using `plot(Result2PL)`

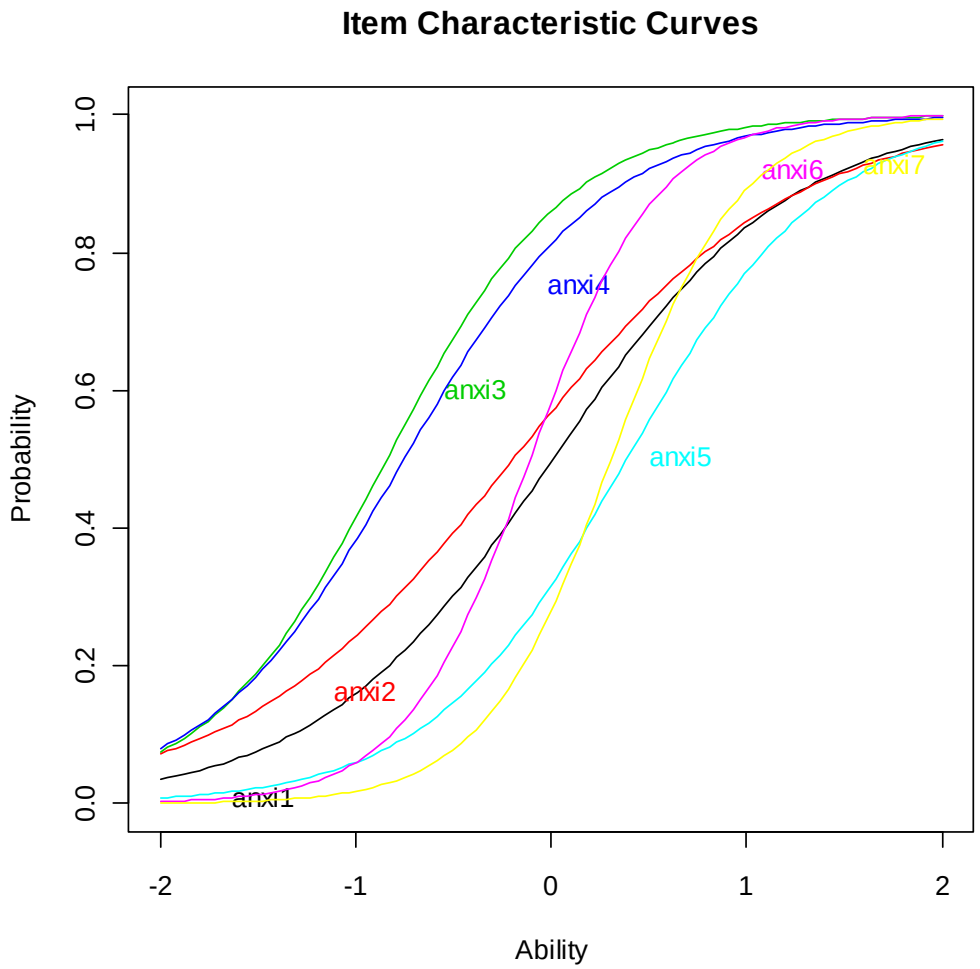
```
plot(x, type = c("ICC", "IIC"), items = NULL,
     zrange = c(-3.8, 3.8), z = seq(zrange[1], zrange[2], length = 100),
     annot, labels = NULL, legend = FALSE, cx = "topleft", cy = NULL,
     ncol = 1, bty = "n", col = palette(), lty = 1, pch, xlab, ylab,
     main, sub = NULL, cex = par("cex"), cex.lab = par("cex.lab"),
     cex.main = par("cex.main"), cex.sub = par("cex.sub"),
     cex.axis = par("cex.axis"), plot = TRUE, ...)
```

Looking at estimates from 2PL



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

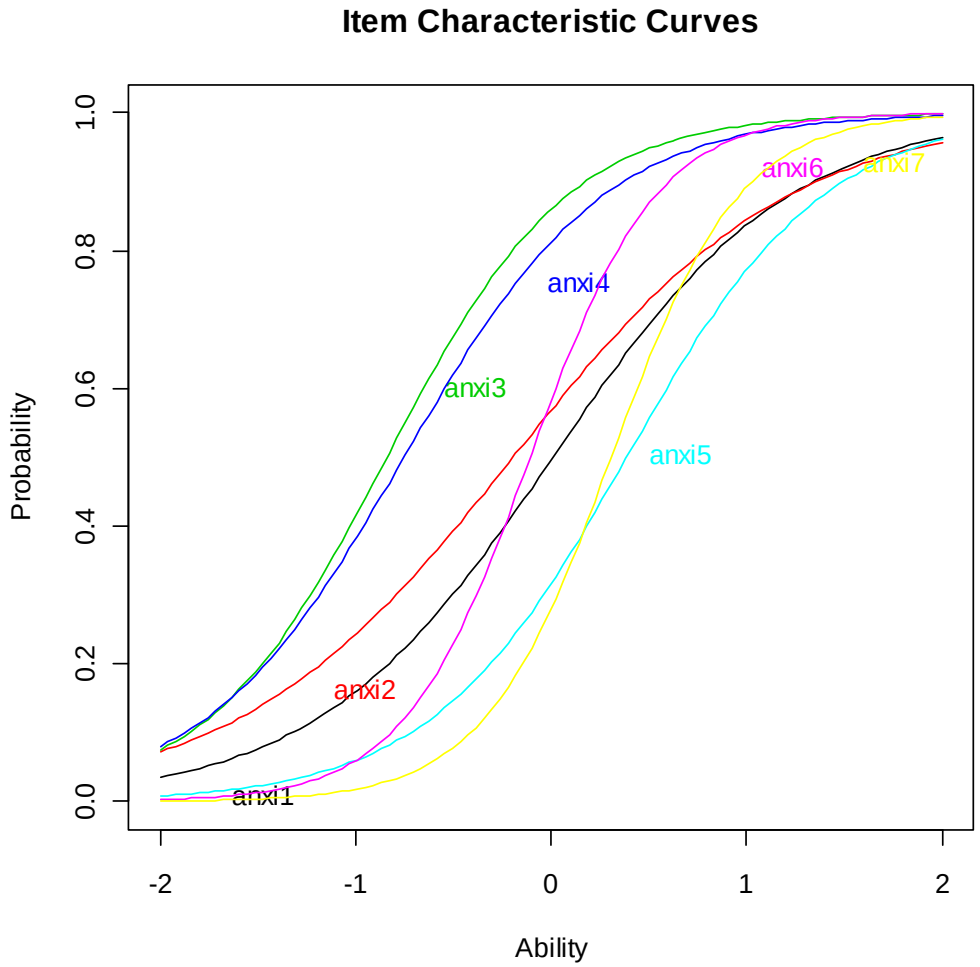
Looking at estimates from 2PL



- items still differ in their difficulty:
  - easiest (b = -.84): anxi3 / GHQ16 "constant strain"
  - most difficult (b = .39): anxi5 / GHQ19 "panicky / scared"

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

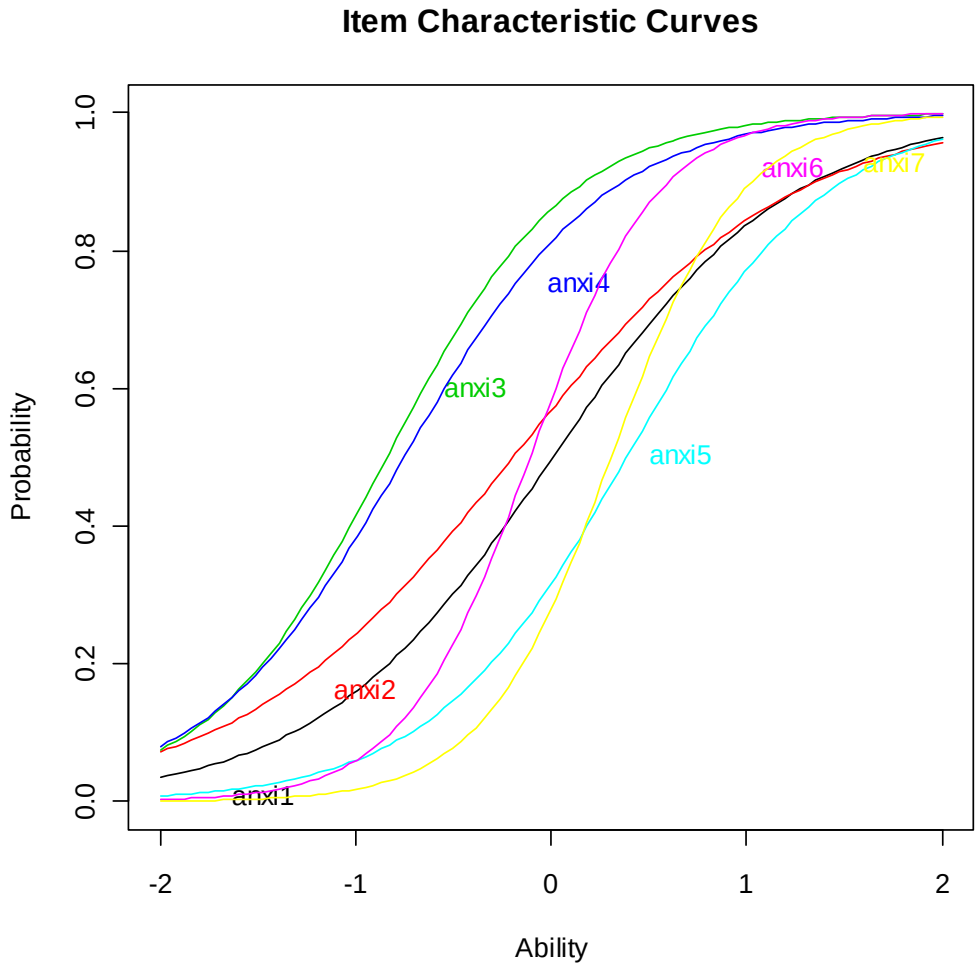
Looking at estimates from 2PL



- 2PL estimates also a for every item,
- providing information on how strongly items discriminate between high and low values on the trait

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Looking at estimates from 2PL



- in this case:
  - highest discrimination (a = 3.10): anxi6 / GHQ20 "everything on top of me"
  - lowest (a = 1.42): anxi2 / GHQ09 "difficulty staying asleep"

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Testing 2PL

- Question: does the model describe the data sufficiently?
- the Pearson-χ^2 bootstrapped test not possible with 2PL
- but we can use the dimensionality test again

Testing 2PL: dimensionality

- Result in our case:

```
Dimen2PL<-unidimTest(Result2PL,Anxiety,B=499)
```

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: 0.5279

Average of second eigenvalues in Monte Carlo samples: 0.3416

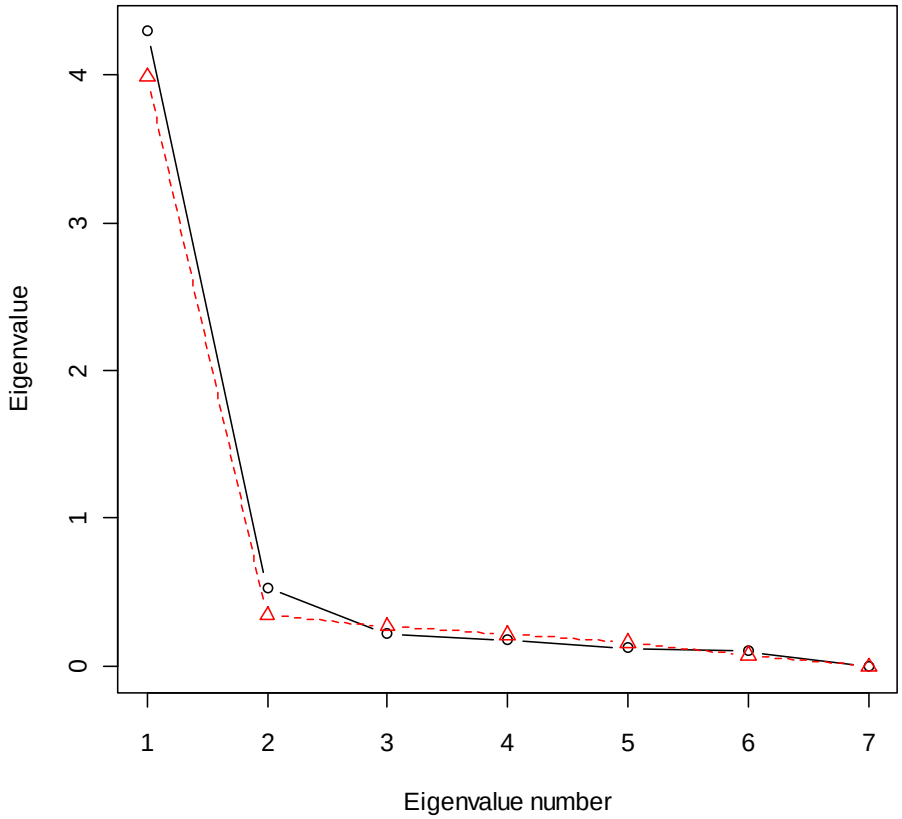
Monte Carlo samples: 499

p-value: 0.002

Testing 2PL: dimensionality

```
plot(Dimen2PL,
   type="b",pch=1:2)

legend("topright",
   c("Real Data",
   "Average Simulated
   Data"), lty = 1,pch =
   1:2, col = 1:2, bty =
   "n")
```



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Adding complexity II: The 3PL Model

ARE GUESSING OR OTHER BASE-RATES INVOLVED?

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

To estimate a 3PL model

- this model uses three parameters in addition to the ability:
  - the items differ in their difficulty
  - and also their slopes / discrimination (a)
  - as well as in their base rate / guessing (c)

$$P(\theta_h) = c + (1-c)\frac{e^{a(\theta_h - b)}}{1+e^{a(\theta_h - b)}}$$

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

To estimate a 3PL model

- Go back to package description:

```
tpm(data, type = c("latent.trait", "rasch"), constraint = NULL,
    max.guessing = 1, IRT.param = TRUE, start.val = NULL,
    na.action = NULL, control = list())
```

- "tpm()" is the function that will estimate the 3PL model (with/without constraints)
- the above provides the structure of the command with all options

To estimate a 3PL model

- for an unconstrained 3PL model only the data matrix has to be specified on which it should be calculated:

```
Result3PL<-
  tpm(Anxiety,control=c(iter.qN=10000,
  GHk=50,verbose=TRUE),
  start.val="random")
```

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

To estimate a 3PL model

```
summary(Result3PL)

Model Summary:
   log.Lik       AIC       BIC
 -10917.87 21877.74 22003.17

Coefficients:
                value std.err   z.vals
Gussng.anxi1   0.0009  0.0051   0.1669
Gussng.anxi2   0.0060  0.0336   0.1776
Gussng.anxi3   0.0031  0.0275   0.1141
Gussng.anxi4   0.0870  0.0878   0.9908
Gussng.anxi5   0.0008  0.0042   0.2024
Gussng.anxi6   0.0079  0.0285   0.2785
Gussng.anxi7   0.0098  0.0086   1.1390
```

Guessing parameters: all near to zero

To estimate a 3PL model

```
summary(Result3PL)

Dffclt.anxi1  0.0083  0.0345   0.2393
Dffclt.anxi2 -0.1847  0.0609  -3.0358
Dffclt.anxi3 -0.8326  0.0831 -10.0162
Dffclt.anxi4 -0.6235  0.0656  -9.5023
Dffclt.anxi5  0.3863  0.0472   8.1795
Dffclt.anxi6 -0.0976  0.0365  -2.6730
Dffclt.anxi7  0.3197  0.0648   4.9336

Dscrmn.anxi1  1.6567  0.0858  19.3063
Dscrmn.anxi2  1.4327  0.1004  14.2679
Dscrmn.anxi3  2.1986  0.1364  16.1127
Dscrmn.anxi4  2.1553  0.2561   8.4142
Dscrmn.anxi5  1.9926  0.1067  18.6727
Dscrmn.anxi6  3.1759  0.3534   8.9869
Dscrmn.anxi7  3.2714  0.2776  11.7828
```

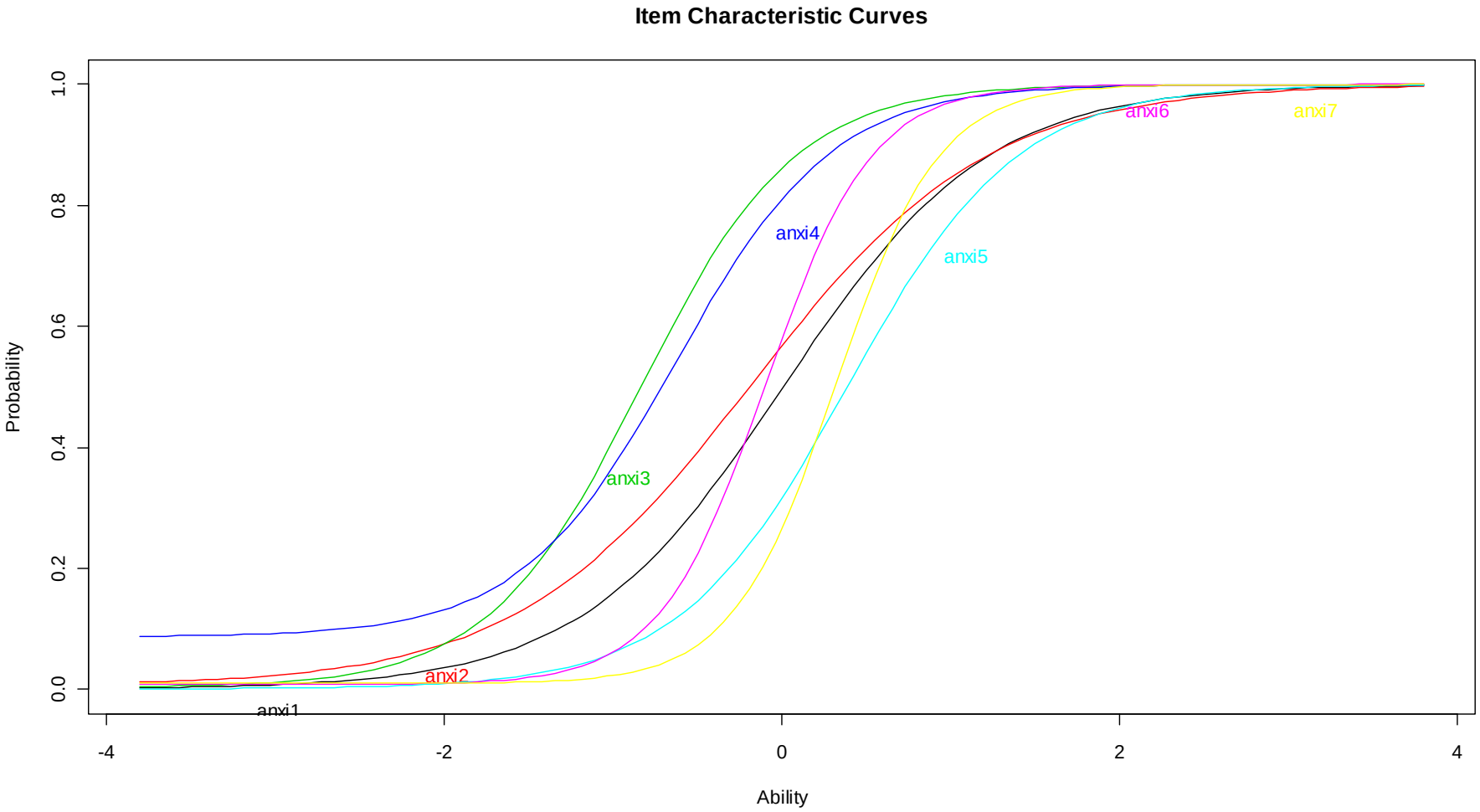
Difficulties and discriminations nearly unchanged compared with 2PL model estimates

Looking at estimates from 3PL

- results from the estimation can also directly plotted by using `plot(Result3PL)`

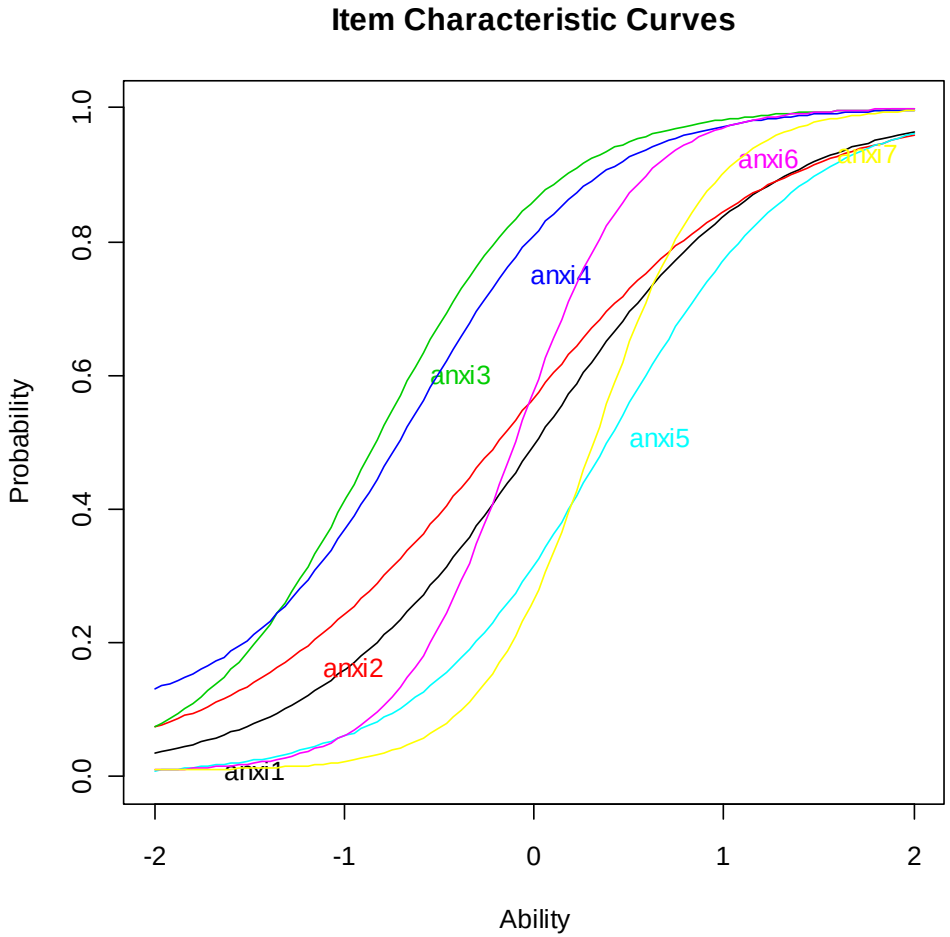
```
plot(x, type = c("ICC", "IIC"), items = NULL,
     zrange = c(-3.8, 3.8), z = seq(zrange[1], zrange[2], length = 100),
     annot, labels = NULL, legend = FALSE, cx = "topleft", cy = NULL,
     ncol = 1, bty = "n", col = palette(), lty = 1, pch, xlab, ylab,
     main, sub = NULL, cex = par("cex"), cex.lab = par("cex.lab"),
     cex.main = par("cex.main"), cex.sub = par("cex.sub"),
     cex.axis = par("cex.axis"), plot = TRUE, ...)
```

Looking at estimates from 3PL



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

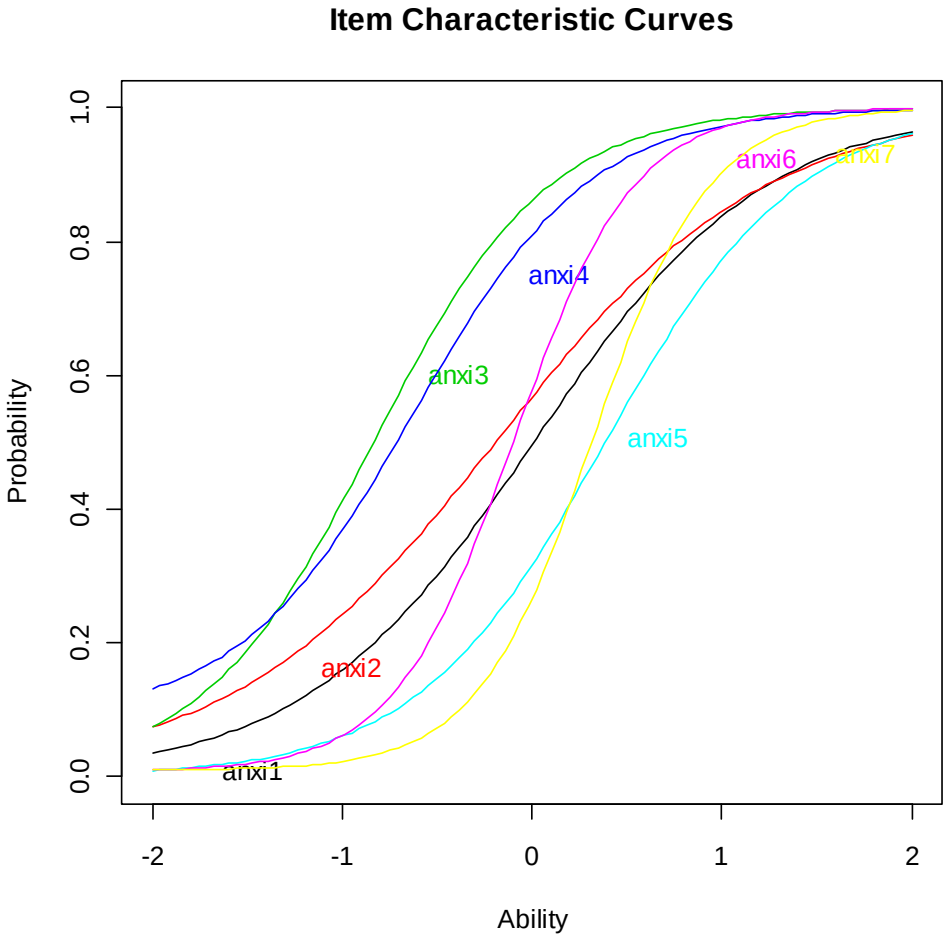
Looking at estimates from 3PL



- items still differ in their difficulty:
  - easiest (b = -.83): anxi3 / GHQ16 "constant strain"
  - most difficult (b = .38): anxi5 / GHQ19 "panicky / scared"

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

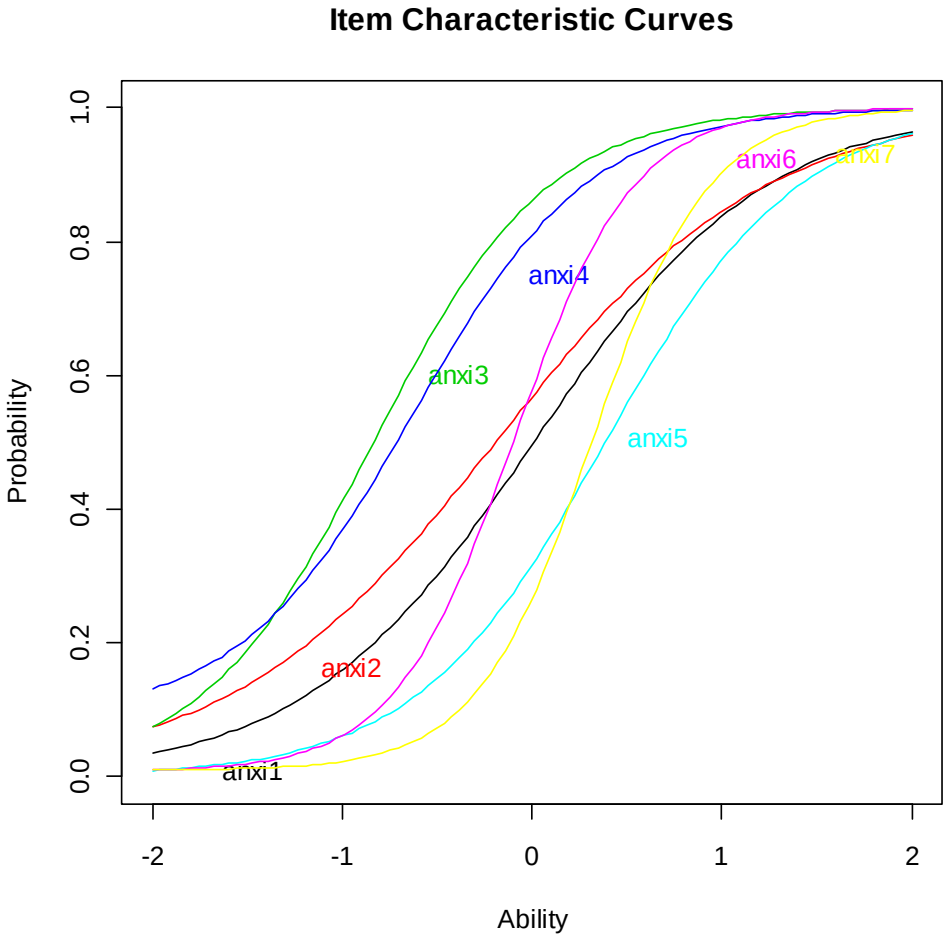
Looking at estimates from 3PL



- discriminations in this case:
  - highest discrimination (a = 3.27): anxi7 / GHQ23 "nervous"
  - lowest (a = 1.43): anxi2 / GHQ09 "difficulty staying asleep"

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

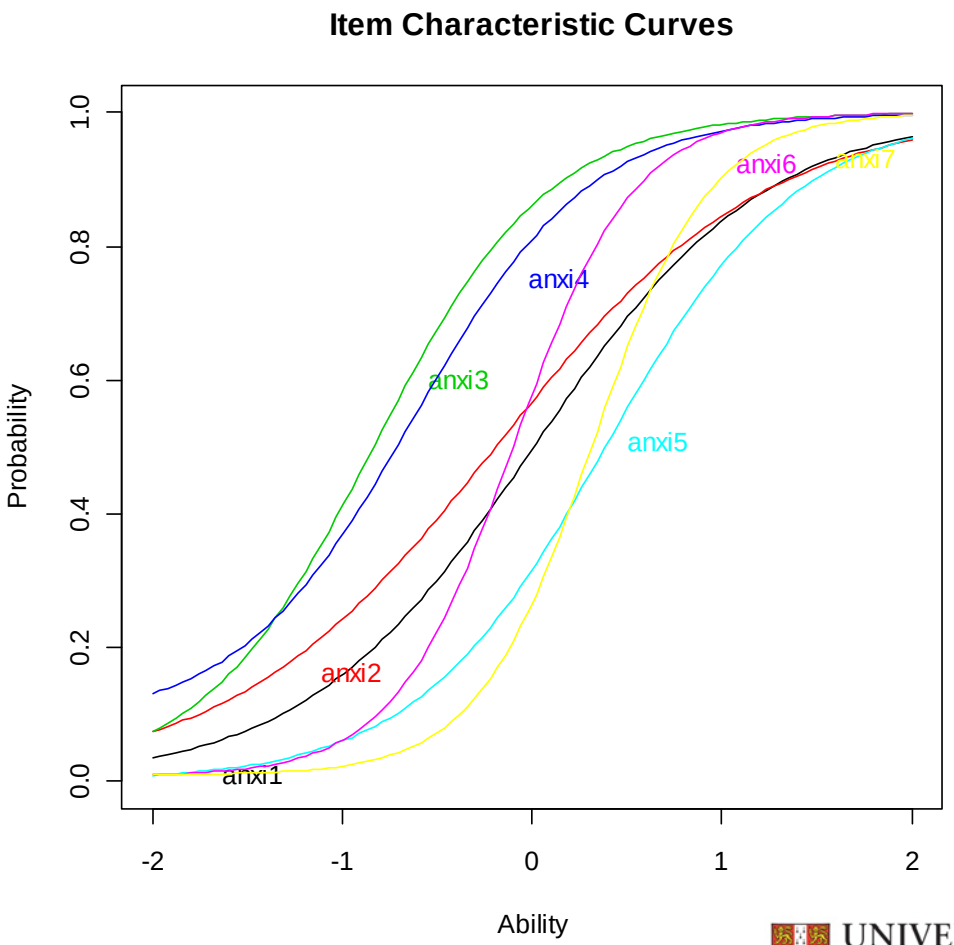
Looking at estimates from 3PL



- 3PL estimates also "c" for every item,
- providing the probability of solving the item with virtually no ability at all ("guessing")

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Looking at estimates from 3PL



- nearly no guessing in this case
- if at all (c = .09) at anxi4 / GHQ18 "edgy"
- i.e.: even with no anxiety-type distress at all, you have a probability of .10 to state that you felt "edgy"

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Testing 3PL: dimensionality

- Result in our case:

```
Dimen3PL<-unidimTest(Result3PL,Anxiety,B=499)
```

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: 0.5279

Average of second eigenvalues in Monte Carlo samples: 0.3466

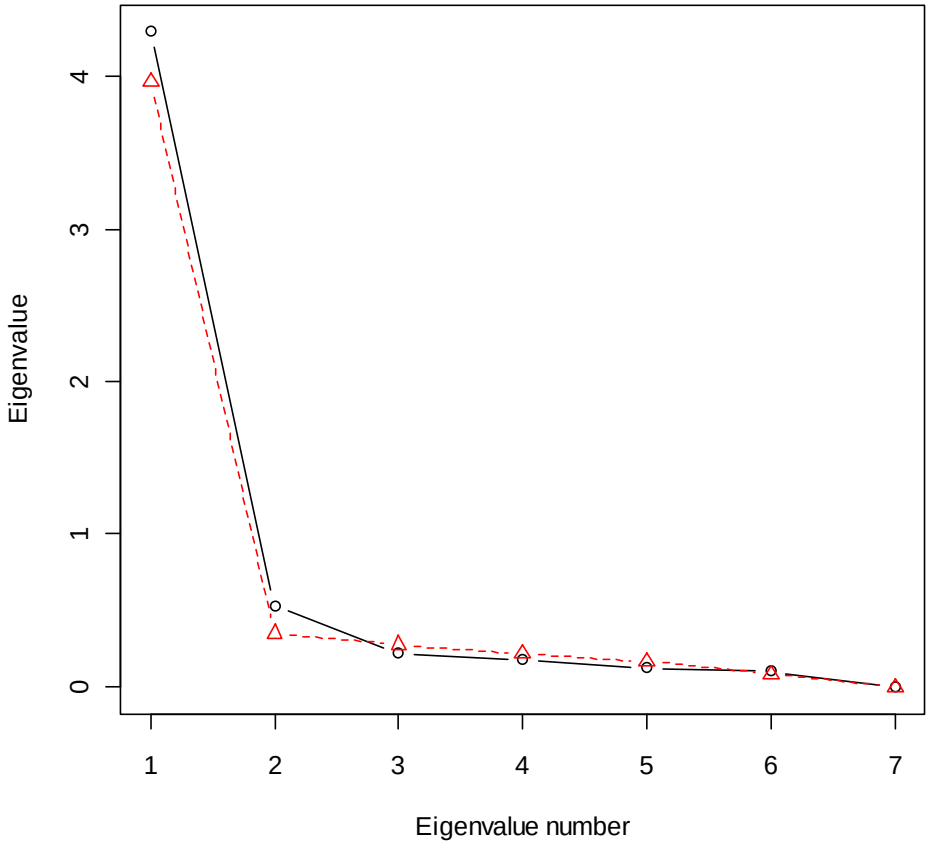
Monte Carlo samples: 499

p-value: 0.002

Testing 3PL: dimensionality

```
plot(Dimen3PL,
   type="b",pch=1:2)

legend("topright",
   c("Real Data",
   "Average Simulated
   Data"), lty = 1,pch =
   1:2, col = 1:2, bty =
   "n")
```



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Practical

- Do the same for one of the other scales!

Comparing models

- Absolute fit tests suffer all from the problem that in most cases the distributional assumptions are not met
- several possibilities exist to look for additional evidence whether the model provides an acceptable fit

Information criteria

- One possibility for models using the same data matrix (but that are not necessarily nested) is to use information criteria to test for parsimony
- Information criteria assess how good the model fits the data (Likelihood) and penalize this fit with the number of parameters needed to obtain this fit

Information criteria

- two typical criteria are AIC and BIC
- Log denoting ln() and n(P) the number of parameters needed

$$AIC = 2\,n(P) - 2LogLike(X)$$

$$BIC = Log(N)\,n(P) - 2LogLike(X)$$

	1PL	2PL	3PL
LogLike	-11002	-10919	-10917
AIC	22020	21867	21877
BIC	22068	21950	22003

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Information criteria

- in this case according to both criteria the 2PL shows the comparatively most parsimonious fit
- as indicated by the lowest values

	1PL	2PL	3PL
LogLike	-11002	-10919	-10917
AIC	22020	21867	21877
BIC	22068	21950	22003

Model test

- Two nested models can also be compared directly by testing the significance in change in likelihood
- "nested": if of two models one of them can be reduced to the other one by fixing parameters, e.g.:
  - 2PL fixing discriminations => 1PL
  - 3PL fixing guessing => 2PL

Model test

- the test statistic for this is calculated as:

$$T = 2\left|LogLike(X_1) - LogLike(X_2)\right|$$

- and it is distributed with the difference in parameters between the two compared models as degrees of freedom

Model test

- Comparison 1PL and 2PL:

```
anova(Result1PL,Result2PL)

 Likelihood Ratio Table
               AIC      BIC   log.Lik    LRT df p.value
Result1PL 22020.99 22068.77 -11002.49
Result2PL 21867.22 21950.84 -10919.61 165.77  6  <0.001
```

The significant statistic indicates that the 2PL model improves the description of the data compared to the 1PL model

Model test

- Comparison 2PL and 3PL:

```
anova(Result2PL,Result3PL)

 Likelihood Ratio Table
               AIC      BIC   log.Lik  LRT df p.value
Result2PL 21867.22 21950.84 -10919.61
Result3PL 21877.74 22003.17 -10917.87 3.48  7   0.838
```

The non-significant statistic indicates that the 3PL model is no improvement compared with the 2PL model

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Item fit

- Measures of person and item fit try to convey a picture of how severe observations deviate from a given model
- since "models are nothing more than an approximation to reality" (Fayers & Machin, 2007: 169) these indices can help further to explore the fit of the model and specific deviations

Item fit

- usually the INFIT / OUTFIT statistics are used which are χ^2-based
- since these are also affected by distributional assumptions, sample size etc.: bootstrap again

Item fit: 1PL and 2PL

(B = 499)

```
Item1PL.fit<-
   item.fit(Result1PL,
   simulate.p.value=TRUE,
   B=9)
```

```
Item1PL.fit
```

	X^2	Pr(>X^2)
anxi1	10.7099	0.996
anxi2	7.6603	1
anxi3	61.9678	0.28
anxi4	32.8981	0.884
anxi5	13.2821	0.928
anxi6	121.1736	0.014
anxi7	91.3611	0.074

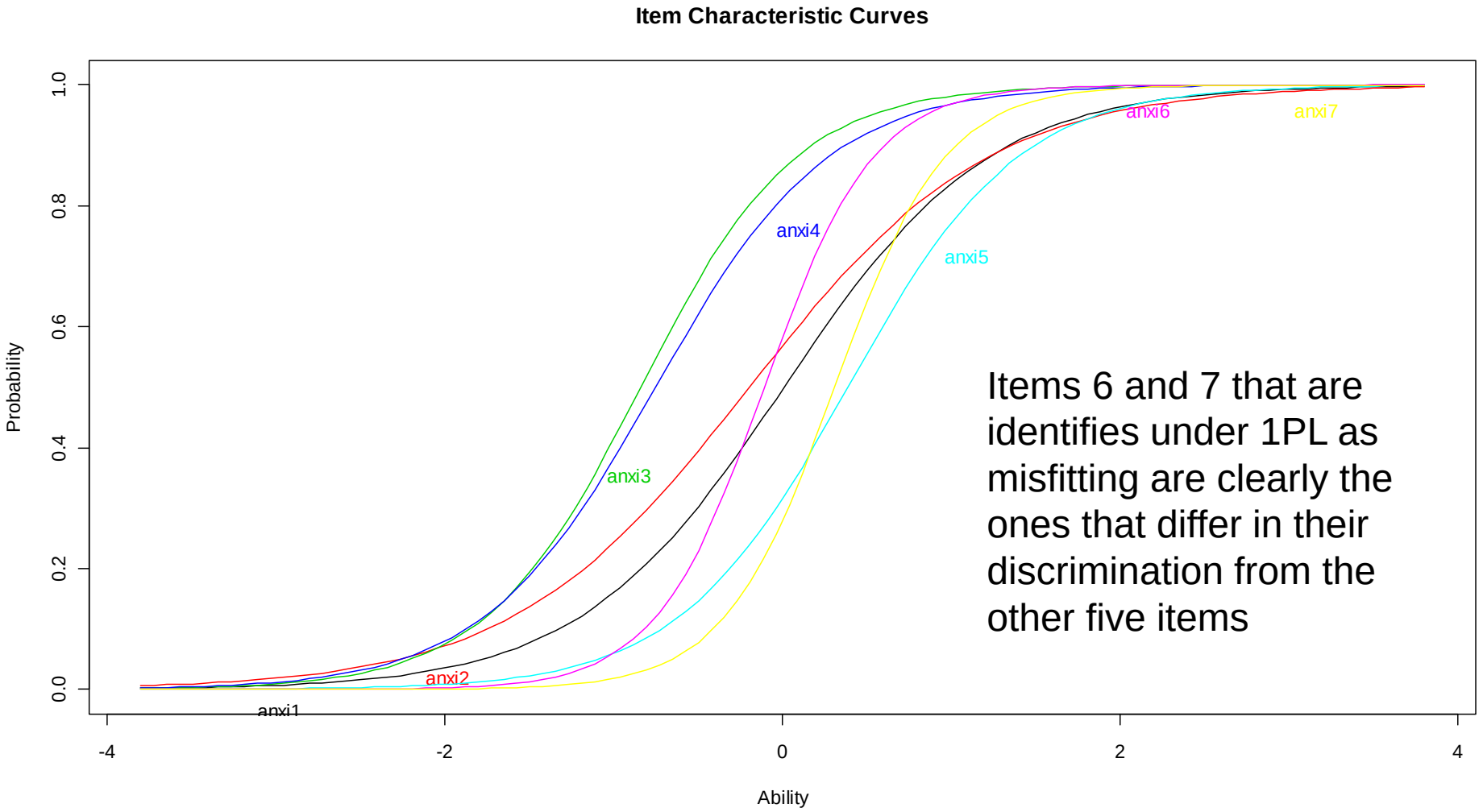
```
Item2PL.fit<-
   item.fit(Result2PL,
   simulate.p.value=TRUE,
   B=499)
```

```
Item2PL.fit
```

	X^2	Pr(>X^2)
anxi1	97.5891	0.886
anxi2	130.6551	0.682
anxi3	254.9275	0.4
anxi4	191.2950	0.256
anxi5	157.6172	0.382
anxi6	116.3733	0.928
anxi7	162.6349	0.36

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Looking at estimates from 2PL



Items 6 and 7 that are identifies under 1PL as misfitting are clearly the ones that differ in their discrimination from the other five items

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Person fit

- Person fit statistics determine whether a response vector of a person significantly deviates from the predicted responses based on that persons estimated ability
- person can be identified to have responses too far off as well as too near to the prediction
- for model fit usually the "off" responses are investigated

Person fit

- test for every person the hypothesis that the response is not predicted by the model:

```
Person1PL.fit<-
  person.fit(Result1PL,alternative="less",
  resp.patterns=Anxiety,simulate.p.value=TRUE,
  B=499)

Person2PL.fit<-
  person.fit(Result2PL,alternative="less",
  resp.patterns=Anxiety,simulate.p.value=TRUE,B
  =499)
```

Person fit

- under H0 about 5% responses would be expected that are not according to the model

```
check1PL<-
  ifelse(Person1PL.fit$p.values<=.05,1,0)
mean(check1PL)
result: 0.011
```

```
check2PL<-
  ifelse(Person2PL.fit$p.values<=.05,1,0)
mean(check2PL)
result: 0.012
```

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Item parameters & Design

- Item parameters show additional perspectives on test design with IRT
- The difficulty parameters show already that instead of developing similar items (like often happens in CTT) the items should cover different areas of the latent trait
- Interpretation of item discrimination is more difficult:
  - high discriminations provide high measurement precision at the specific point on the latent trait (but nowhere else)
  - items with lower discriminations help in a 2PL test to distinguish between the low and high end of the trait
  - items with very low discriminations most likely not related to the latent trait

Final Decision

- 3PL seems not to be justified – simply alone by the non-existing guessing effects on the items
- 2PL seems to be considerable improvement compared to 1PL in terms of precision
- but concerning fit statistics 1PL and 2PL seem to be similar (Item fit, Person fit)

Sample Sizes

- Fayers & Machin (2007, p. 172): "Use large datasets!"
- DeMars (2010):
  - 1PL starting from 100-200 examinees
  - 2PL/3PL: a-parameters well behaved from about 500 examinees on
  - 3PL: accuracy for the c-parameters increases markedly to the level of 2000 examiness and slightly less to about 4000

Sample Sizes

- DeMars (2010) on polytomous:
  - GPC-Model: about 150 for three categories, maybe 1000 or more for six categories
  - Rules of thumb:
    - under very favourable conditions, 20 items: maybe 500
    - less favourable, 40 items: maybe 1000
    - more items, even less favourable 2000 and more

Why IRT?

- assumptions of IRT models are testable in comparison to CTT
- provides information on functioning of single items as well as on the whole group
- provides measurement on higher than interval scale level