UNIVERSITY OF CAMBRIDGE

The Psychometrics Centre

Summer School in Applied Psychometric Principles

Peterhouse College

13th to 17th September 2010

The Rasch Model

Day 3

Jan R. Böhnke
University of Trier, Germany

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Topics already covered

- We have…
  - Introduced IRT
  - Introduced simple models for binary responses
  - Discussed IRT assumptions
  - Introduced models for polytomous responses
  - Discussed assessment of fit for these models

Today

- We will spend a day with the Rasch Model

- Why that?
  - Rasch Model is a very simple test model
  - which has extraordinary measurement qualities
  - can be generalized to several applications
  - and which is testable

The Rasch Model

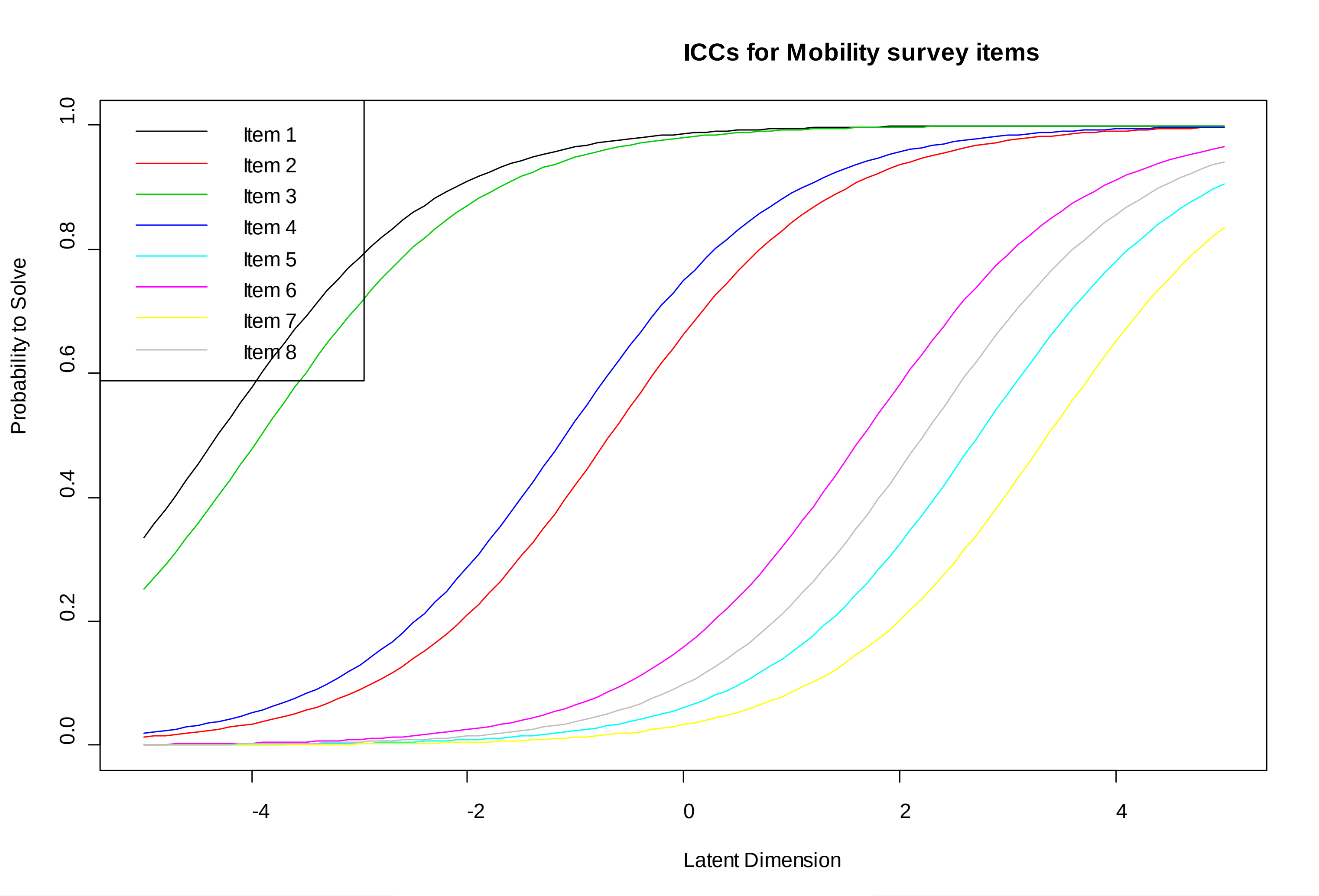
- The Rasch model can be seen as a very reduced / restricted version of the models we already encountered in the course:
  - the slopes for all items are constrained to be equal (usually Dα = 1)
  - no guessing parameter (c = 0)

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

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

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

The Rasch Model



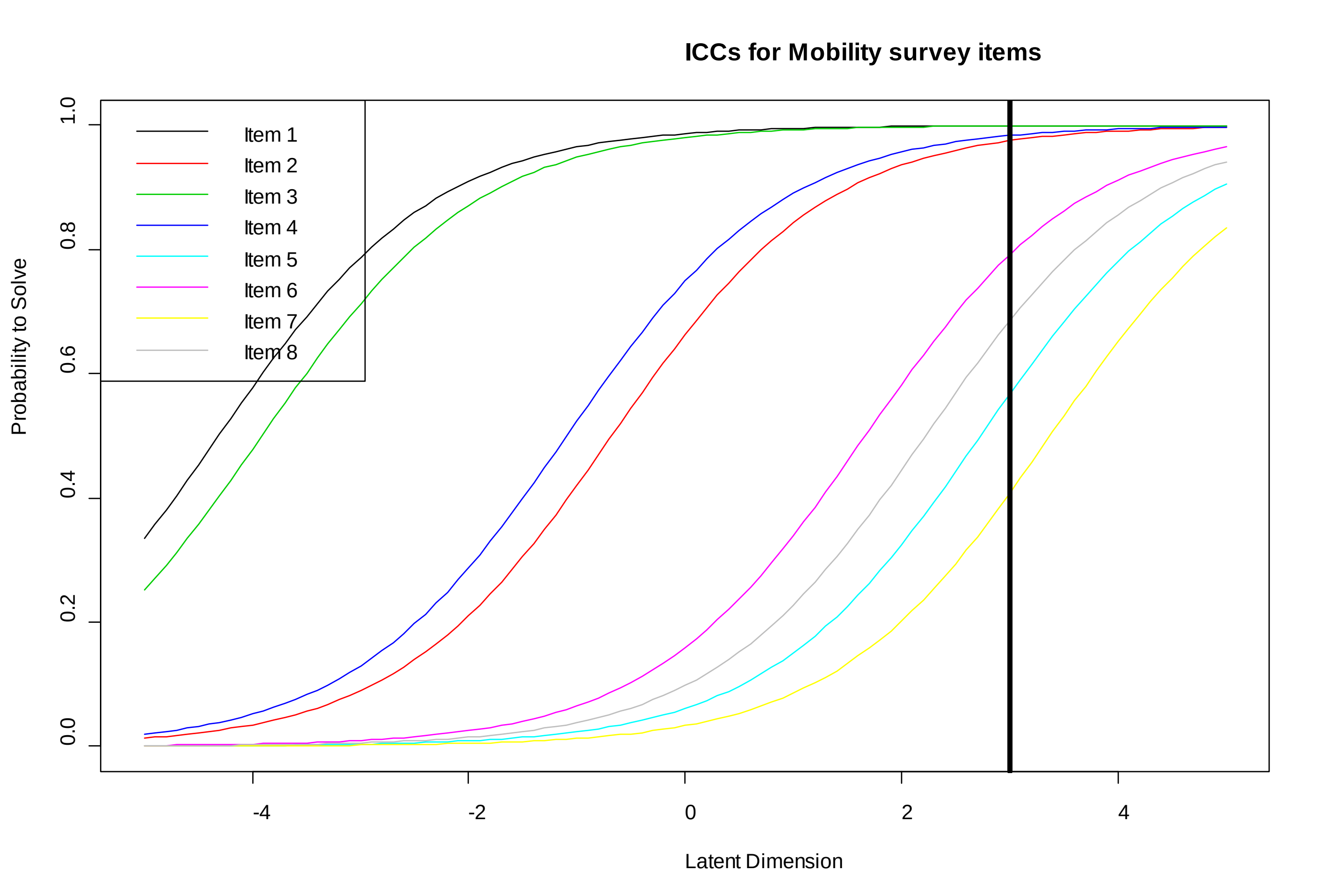
UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

The Rasch Model

- The fact that only one parameter is modeled leads to the models‘ most important consequence:
  - the ICCs are non-intersecting
  - thereby holds for any comparison of persons or items:

$$P(x_{vi} = 1) > P(x_{wi} = 0) \Leftrightarrow \theta_v > \theta_w$$

The Rasch Model



$$P(x_{vi}=1) > P(x_{wi}=0) \Leftrightarrow \theta_v > \theta_w$$

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

The Rasch Model

- This feature of the Rasch Model is called „specific objectivity“; when the Rasch model holds:
  - irrespective of which combination of items from a scale, the same ordering of persons is obtained
  - irrespective of what subsample of persons, the items are ordered the same way according to their difficulty

$$P(x_{vi}=1) > P(x_{wi}=0) \Leftrightarrow \theta_v > \theta_w$$

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

The Rasch Model

- because both these orderings are stable (within measurement error):
  - it is not important which combination of items was solved by a respondent;
  - and from that follows that the sum of solved items contains all information about the respondent‘s position on the latent trait

$$P(x_{vi} = 1) > P(x_{wi} = 0) \Leftrightarrow \theta_v > \theta_w$$

The Rasch Model

- this principle of „specific objectivity“ provides the possibility to construct two specific tests that test whether the data is Rasch-scalable or not:
  - the Andersen Likelihood Ratio Test: checks whether the invariance of item parameters in different subpopulation holds
  - the Martin Löf Test: checks whether the person parameters are invariant by splitting the scale into different subsets of items

Sideline: Guttman Scaling

- In Guttman scaling only specific patterns allowed:
  - items ordered according to their difficulty
  - a person solving a more difficult item has to solve all items that are easier than that

Items			
0	0	0	0
1	0	0	0
1	1	0	0
1	1	1	0
1	1	1	1

Sideline: Guttman Scaling

- „deterministic model“
- only ordinal measurement possible but score also represents all available information on respondents
- Measurement Theorem:

$$(x_{vi} = 1) \wedge (x_{wi} = 0) \Leftrightarrow \theta_v > \theta_w$$

Items

0	0	0	0
1	0	0	0
1	1	0	0
1	1	1	0
1	1	1	1

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Guttman Scaling & the Rasch Model

- In essence the Rasch Model does exactly the same:
  – looking for an ordering of the items that describes persons as well as items on the same scale

Items			
0	0	0	0
1	0	0	0
1	1	0	0
1	1	1	0
1	1	1	1

Guttman Scaling & the Rasch Model

- the Rasch model is in a sense completely different:
  - it acknowledges measurement error: Guttman structure would be the ideal pattern, but deviations from that are possible
  - „probabilistic model“

Items

0	0	0	0
1	0	0	0
1	1	0	0
1	1	1	0
1	1	1	1

Guttman Scaling & the Rasch Model

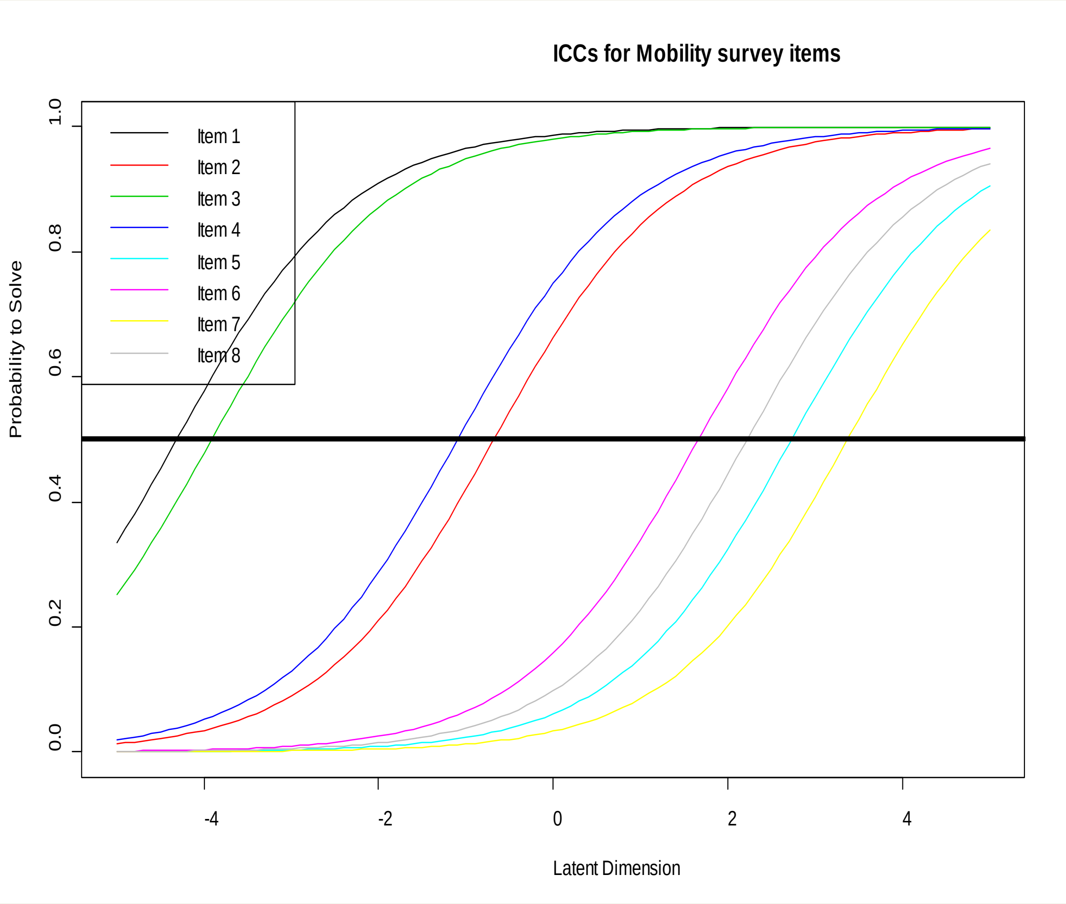
- the Rasch model is in a sense completely different:
  - by introducing a well-behaved mathematical function to describe the relationship between the trait and the probability, it is possible to scale the items and scores on a (more than) interval continuum

Items

0	0	0	0
1	0	0	0
1	1	0	0
1	1	1	0
1	1	1	1

Guttman Scaling & the Rasch Model

- based on the data it can be assessed, where on the latent continuum the item is solved with a probability of 50%
- since the slope is defined by the mathematical function, distances between the locations can be measured



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Dimensionality or Local independence assumption

- Item responses are independent after controlling for (conditional on) the latent trait
- There is only one dimension explaining variance in the item responses
  - based on this assumption non-parametric tests can already be employed to check whether the data fits the model BEFORE we even estimate the model (e.g. Ponocny, I. (2001). *Psychometrika, 66, 437-460*.)

Features of the Rasch-Model

- Two core differences to other IRT models:
  - it can be tested whether the respondents‘ patterns in the answer vectors comply with the assumtion of the Rasch Model (tests not based on „by-proxy“ tests with factor analysis)
  - Compared to the other models the score is the „sufficient statistic“; in the other models it is a weighted sum

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

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)

Estimating item parameters

- Conditional maximum likelihood (CML)

Formulas not important in detail, but:
the estimator for every item parameter depends
a) on the interaction of the location of all other items
b) conditional on all test scores

$$\hat{\delta}_i = \ln\left(\sum_{r=1}^{k} n_r \frac{\gamma_{r-1}^{(i)}(\varepsilon)}{\gamma_r(\varepsilon)} \right) - \ln(x_{oi})$$

$$\ln(CL(X)) = \sum_{i=1}^{k} x_{oi} \ln(\varepsilon_i) - \sum_{r=0}^{k} n_r \ln(\gamma_r(\varepsilon))$$

(Wilhem Kempf, University of Konstanz)

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
- Warm’s Maximum Likelihood (WML)
  - often employed (e.g. WINMIRA) because it provides estimates for full/empty response patterns
  - more computational intensive than ML
  - more central estimates; SEs equal to ML
- Spline interpolation
  - estimator based on the relationship between scores and estimated person parameters
  - employed in eRm

EMPIRICAL EXAMPLE: BDI-DATA

PRACTICAL: MOBILITY DATA

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Practical: Mobility survey

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

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Estimation in R – eRm

library(eRm)

ResMob<-RM(*Itemmatrix*,se=TRUE,sum0=TRUE)

Itemmatrix is the Matrix containing the responses

se=TRUE (standard errors are estimated)

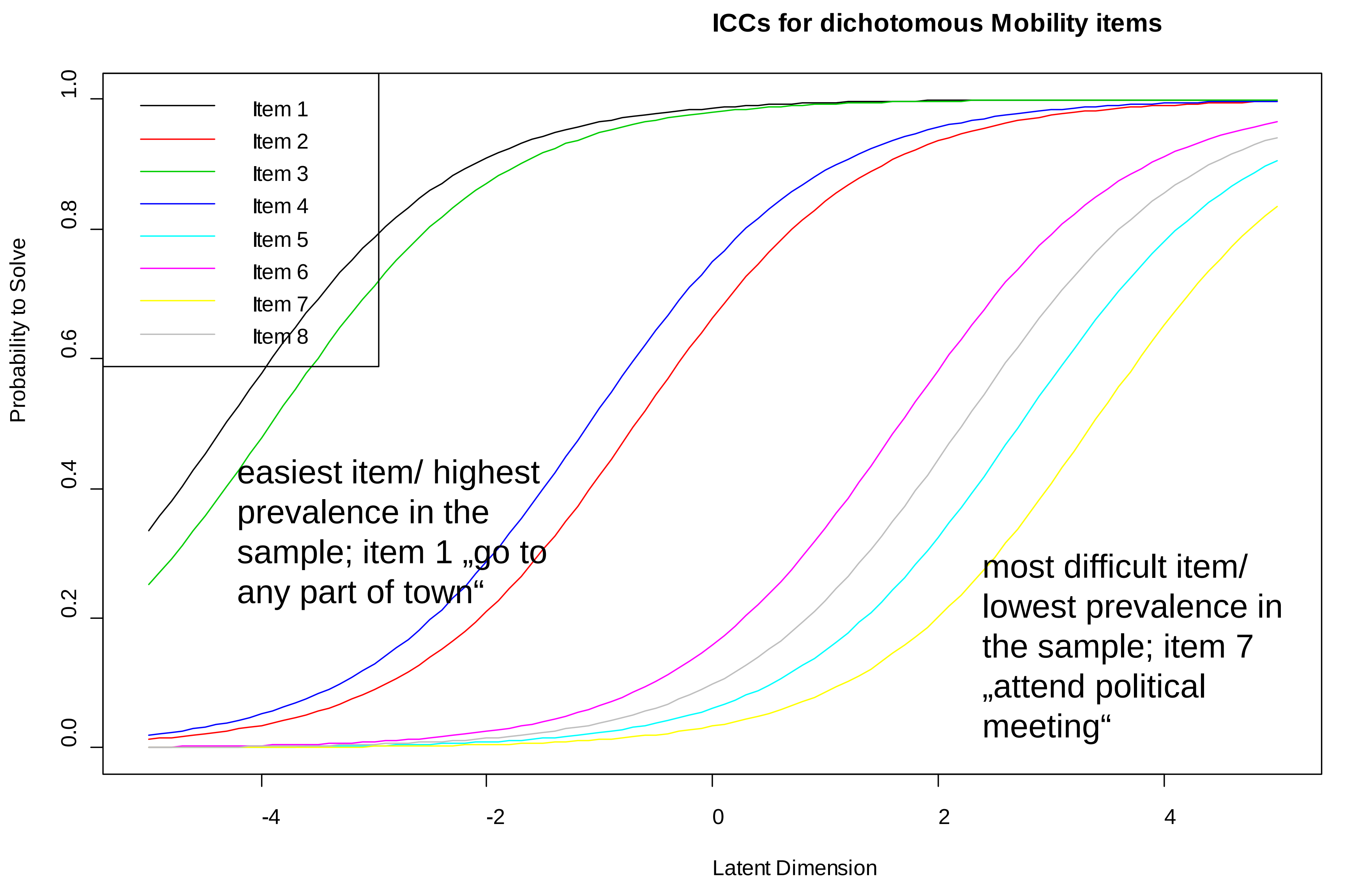
sum0=TRUE (b's are normed on 0)

Plotting

```
plotjointICC(ResMob, main="ICCs for dichotomous Mobility items", xlim=c(-5,5),legpos="topleft")

plotjointICC(ResMob, main="ICCs for dichotomous Mobility items", item.subset=c(1,5,7))
```

University of Cambridge
The Psychometrics Centre

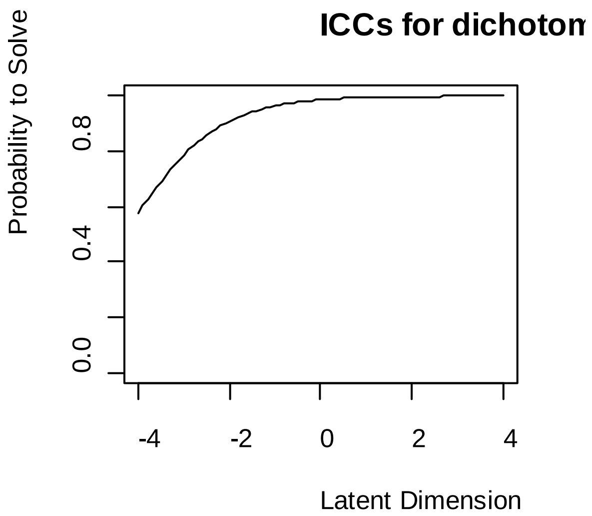
Item Characteristic Curves



Plotting

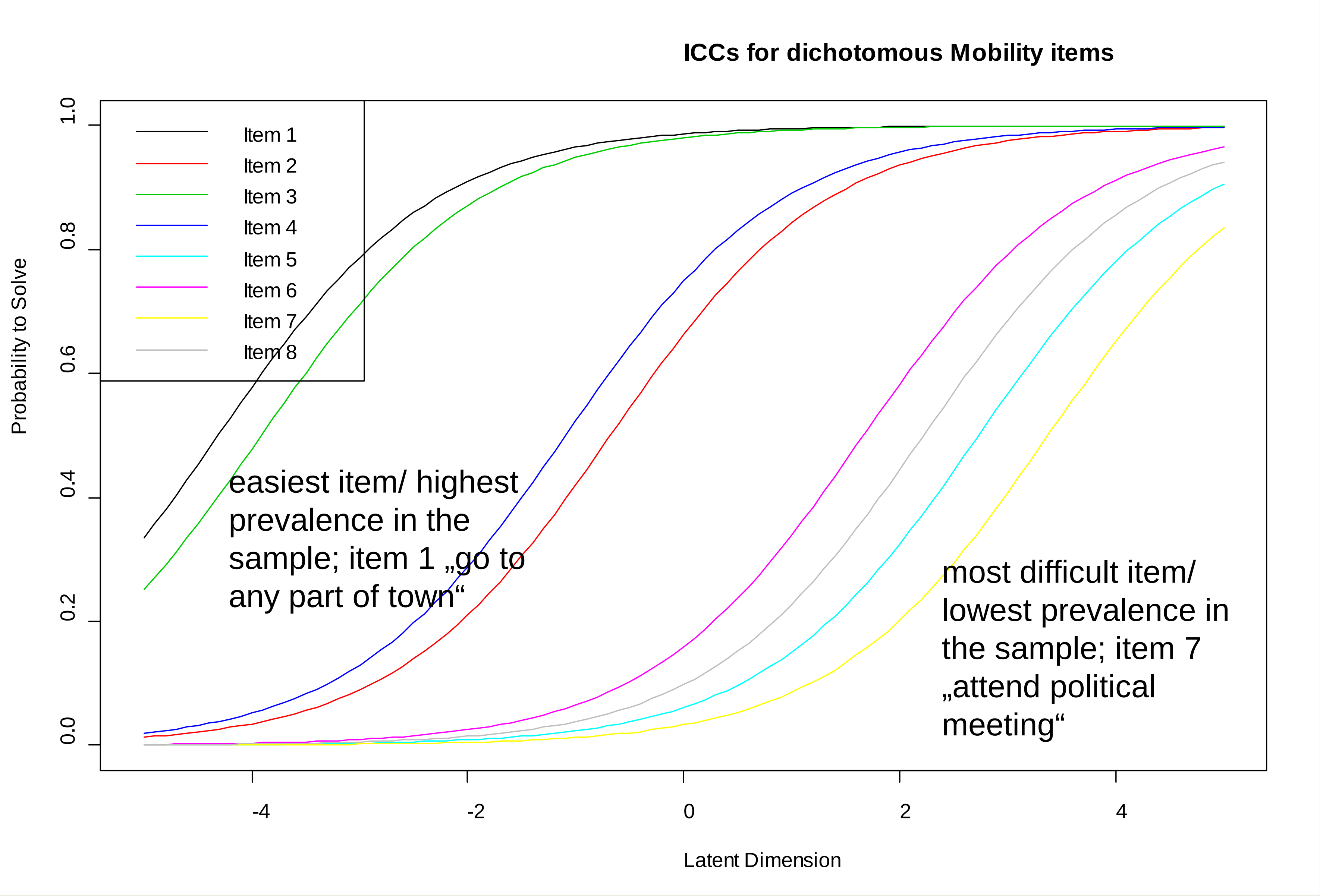
```
plotICC(ResMob,empICC=list("kernel"),empCI=list(),main="ICCs for dichotomous Mobility items")
```

Plotting

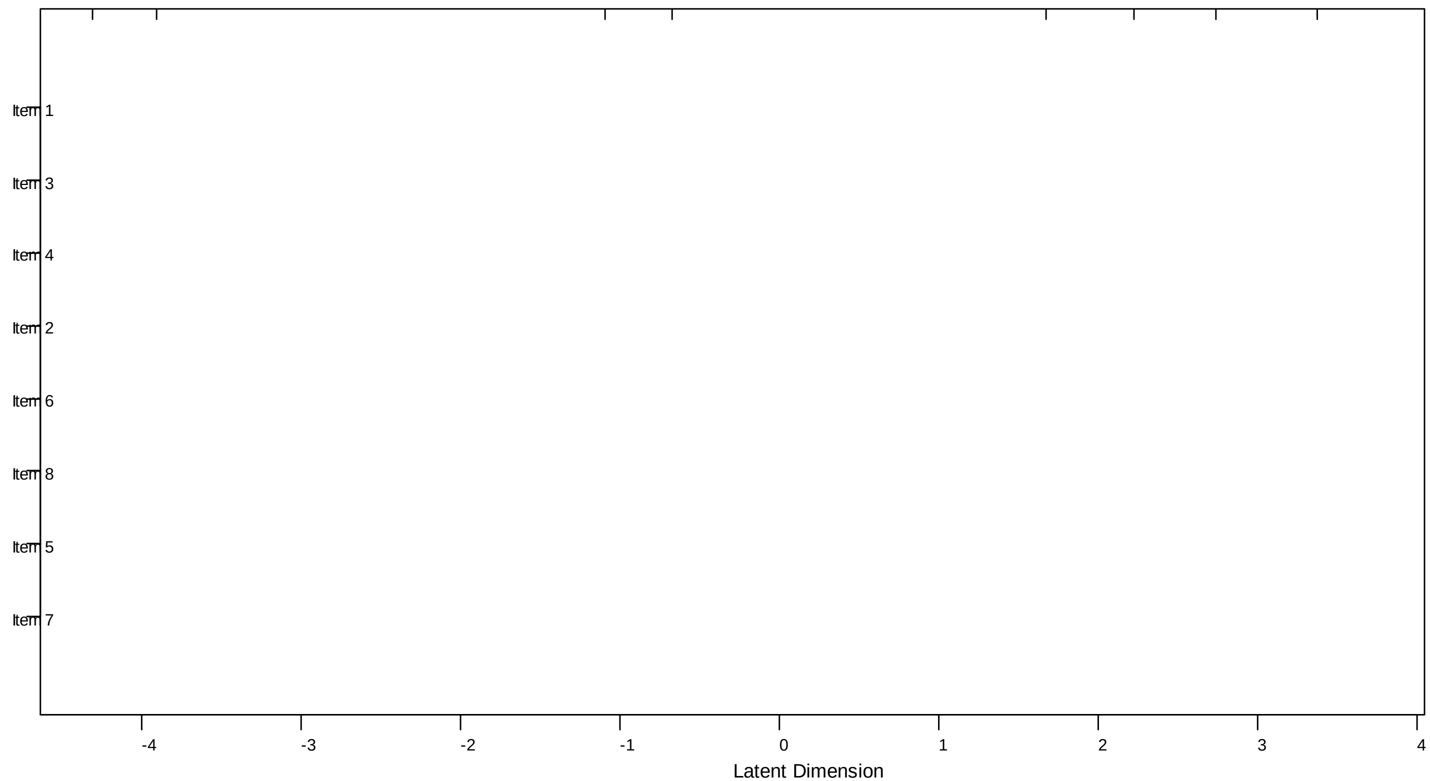


ICCs for dichotom
Probability to Solve
0.0
0.4
0.8
-4
-2
0
2
4
Latent Dimension

Item Characteristic Curves



Joint distribution of items and person parameters

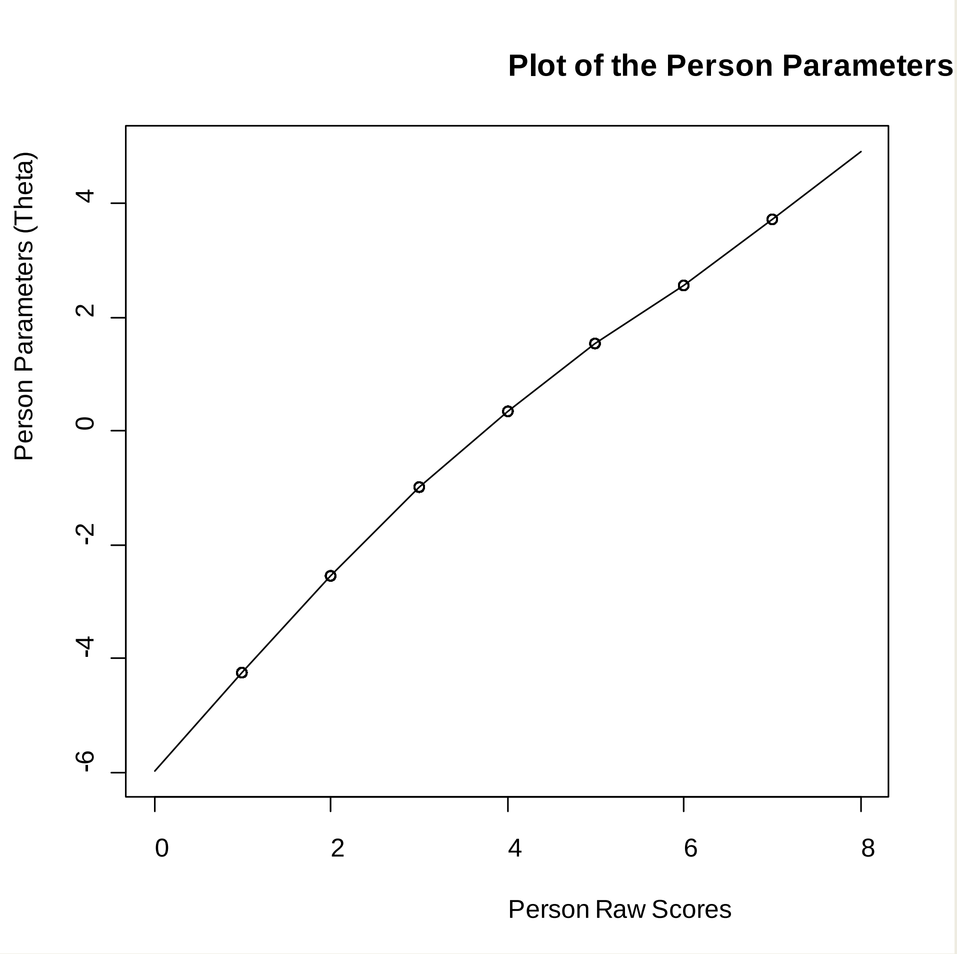


Estimating the person parameters

```
PersMob<-person.parameter(ResMob)
plot(PersMob)
```

Relationship between scores and person parameters

- every score can be transformed into the scale-free metric of the person parameters
- not related in linear fashion (esp. in the tails)
- also: there are only as many person parameters estimated as possible scores (unlike in the other IRT models)



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

What if…?

- What would be won if the Rasch-Model fitted the data?
  - we know that the summed item score can be used as a simple descriptive measure for the ability (was also used to estimate the model)
  - we also would have the person parameters to represent the ability on a (better than) equal interval level
  - we would know that the test is fair at any rate („specific objectivity“)

- The nice thing about the Rasch-Model is, that clear predictions about the nature of the data follow from the model formulation and these predictions can be easily tested

Testing the Rasch Model

- Non-Parametric tests:
  - Ponocny, I. (2001). Psychometrika, 66, 437-460.
  - before estimating the Rasch Model at all we could test whether the observed item responses of the persons would be expected if the test was Rasch scaled
  - not covered in detail here

Testing the Rasch Model

- Parametric Tests based on “specific objectivity”:
  - ANDERSEN‘S LR-TEST: all estimated parameters are independent of the subgroup of the sample in which they are estimated (e.g. gender)
  - MARTIN LÖF-TEST: irrespective of which items are used, the comparison between two test persons should result in the same ordering

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Andersen‘s Likelihood Ratio Test

- Procedure:
  - The Rasch Model is estimated independently in both/all subgroups
  - and then the fit is compared using the likelihood:

$$\chi \mathrm{I} = \text{-}2 * (\mathrm{LN}(\mathrm{Likelihood}(\text{full data set}) + \sum_{g} (\mathrm{LN}(\mathrm{Likelihood}(\mathrm{Subgroup}_{\mathrm{g}}))$$

- with df=(g-1)*(k-1); with g = number of subgroups and k = number of items

- these Likelihoods should be the same, if the itemparameters (δ_i) were the same in all subgroups g, i.e. the test should be non-significant

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Andersen‘s Likelihood Ratio Test

- the default test is with high vs. low scorer groups
- Sample is divided into two groups:
  - a: scores <= median;
  - b: scores > median

- Andersen1<-LRtest(ResMob,se=TRUE)
- summary(Andersen1)

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Andersen's Likelihood Ratio Test

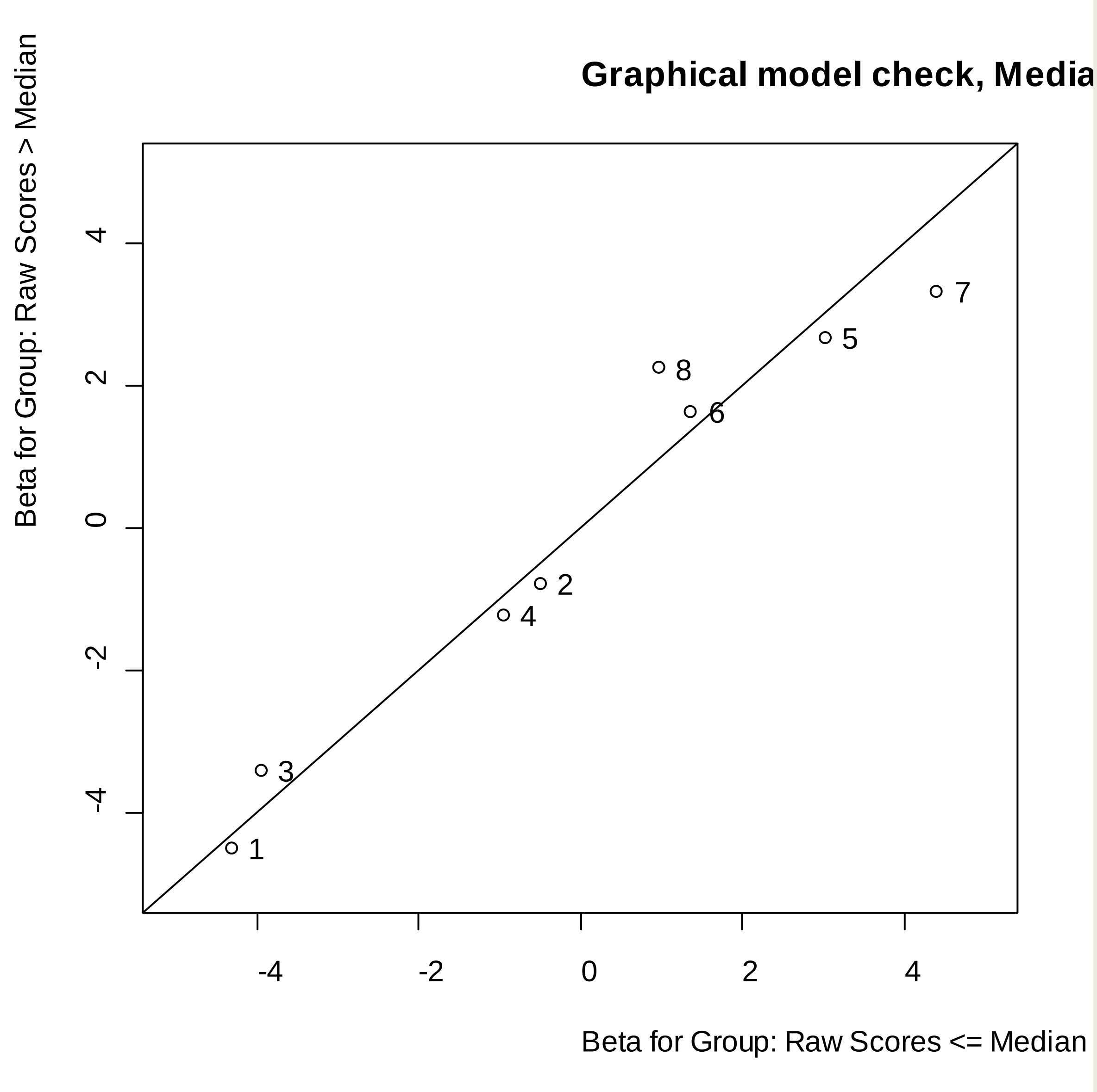
- the default test is with high vs. low scorer groups
- Sample is divided into two groups:
  - a: scores <= median;
  - b: scores > median

- $\chi^2 = 78.36$ with $df=7$; $p < .001$
- the 8 items do not have the same difficulty parameters in both samples

Andersen's Likelihood Ratio Test

- Plotting:

```
plotGOF(Andersen1,main="Graphical model
  check, Median",tlab="number",
  ctrline=list(gamma=0.95, col="blue",
  lty="dashed"), conf=list(),xlim=c(-5,5),ylim=c(-
  5,5))
```



Graphical model check, Media
Beta for Group: Raw Scores > Median
Beta for Group: Raw Scores <= Median
-4
-2
0
2
4
1
2
3
4
5
6
7
8

Andersen‘s Likelihood Ratio Test

- no covariates in the data file; therefore simulate one:

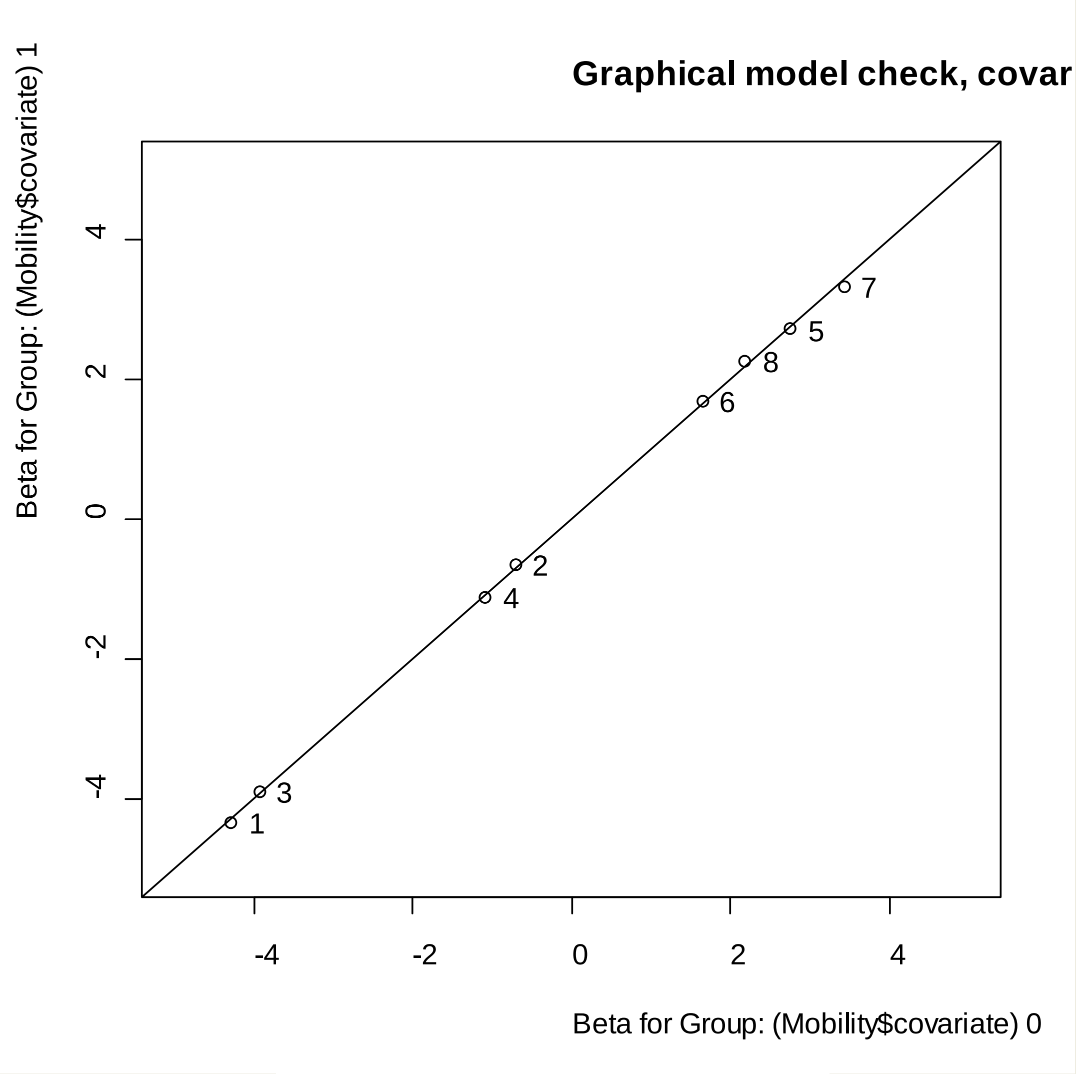
```
Mobility$covariate<-
  with(Mobility,rbinom(8445,1,.5))


Andersen2<-
  LRtest(ResMob,se=TRUE,splitcr=(Mobility$cov
  ariate))
```

Andersen‘s Likelihood Ratio Test

- the random split results in a non-significant test statistic:

- $\chi^2 = 3.15$ with df=7; $p = .87$
- the 8 items do have the same difficulty parameters in both samples

Andersen‘s Likelihood Ratio Test



Wald Test

- Both tests provide only information on the fact that the difference between groups is at least for one item parameter big enough, to produce a significant test statistic
- Wald-Tests can be used to test the differences between the subgroups for every item

$$z_i = \frac{\hat{\beta}_{i1} - \hat{\beta}_{i2}}{\sqrt{{\sigma_{i1}}^2 + {\sigma_{i2}}^2}}$$

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Wald Test

- Syntax for split with median raw score splitting:

- Wald1<-Waldtest(ResMob)

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Wald Test

- In this example done for median of ability
- The following items fail this test:
  - Item 3: $p = .002$
  - Item 8: $< .001$

- typical post-hoc questions apply: Type I error, cross-validation,…

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Differential Item Functioning

- These ideas are closely connected to the question of Differential Item Functioning (DIF)
- DIF explores whether there are systematic differences between groups in the difficulty of endorsing specific item categories
- these should not be present (or corrected for), because they question the fairness of a specific test
- topic of tomorrow

Martin Löf Test

- Procedure:
  - The Rasch Model is estimated independently in both ITEM subgroups
  - then the fit is compared using the likelihood:

$$\chi \mathrm{I} = -2 * (\mathrm{LN}(\mathrm{Likelihood}(\text{full data set}) + \sum_{l} (\mathrm{LN}(\mathrm{Likelihood}(\mathrm{Subgroup}_{\mathrm{l}}))$$

  - For two subgroups with df=(I_1*I_2-1); with I_1 = number of items in subgroup 1 and I_2 = number of items in subgroup 2
  - these Likelihoods should be the same, if the itemparameters (θ_j) were the same in all subgroups, i.e. the test should be non-significant

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Martin Löf Test

- the default test is with items high vs. Low in difficulty
- Sample is devided into two groups:
  - a: itemparameter <= median (Items: 1, 2, 3, 4);
  - b: itemparameter > median (Items: 5, 6, 7, 8);

- χ^2 = ~3438 with df=15; p <<< .001
- The items are (at least with this split criterion) not homogeneous

Martin Löf Test

- Other splits possible, e.g.:
  - One has a hypothesis which items should be grouped together more closely
  - Random splits
- Please think of sub grouping / sub scaling! Then we will perform the test for this specific comparison!

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Assessing Model Fit: Summary

- (Some) Ways to test the fit of the Rasch-Model:
  - Andersen‘s LR-Test: Itemparameters the same for different subgroups?
  - Wald-Tests: Itemparameters the same for different subgroups (pay attention to alpha-level!)
  - Martin-Löf-Test: Personparameters are the same when resulting from different item-sets

Assessing Model Fit: Summary

- Splits in this regard are usually only as good as the observed criteria
- Rost & von Davier (1997) proposed therefore:
  - estimate the Rasch-Model on your data
  - estimate a two class Mixed Rasch Model on the same data to identify the maximal possible differences between persons in response patterns
  - LR-test between these models or (my opinion) Andersen test with these groups

POLYTOMOUS RASCH MODEL

Polytomous Rasch Models

- The question for polytomous IRT models is, how the different categories can be mapped on the latent continuum
- already seen: Graded Response Model
- In the Rasch perspective especially the *Partial Credit Model* is of interest
- and the constraint version of the so-called *Rating Scale Model*

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

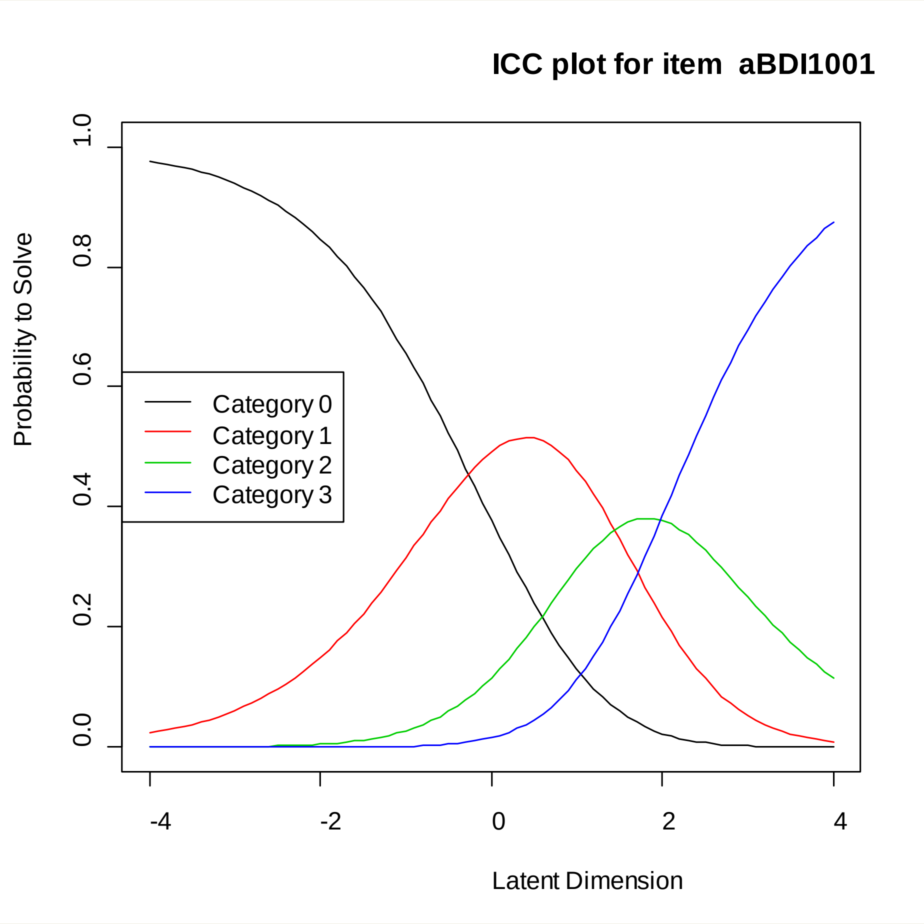
$$P_{i0}(\theta) = \frac{\exp\left[0 + a_i(\theta - b_{i1})\right]}{\exp[0] + \exp\left[a_i(\theta - b_{i1})\right] + \exp\left[0 + a_i(\theta - b_{i1}) + a_i(\theta - b_{i2})\right]}$$

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 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

Interpretation

- 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



Estimating a Rating Scale Model in eRm

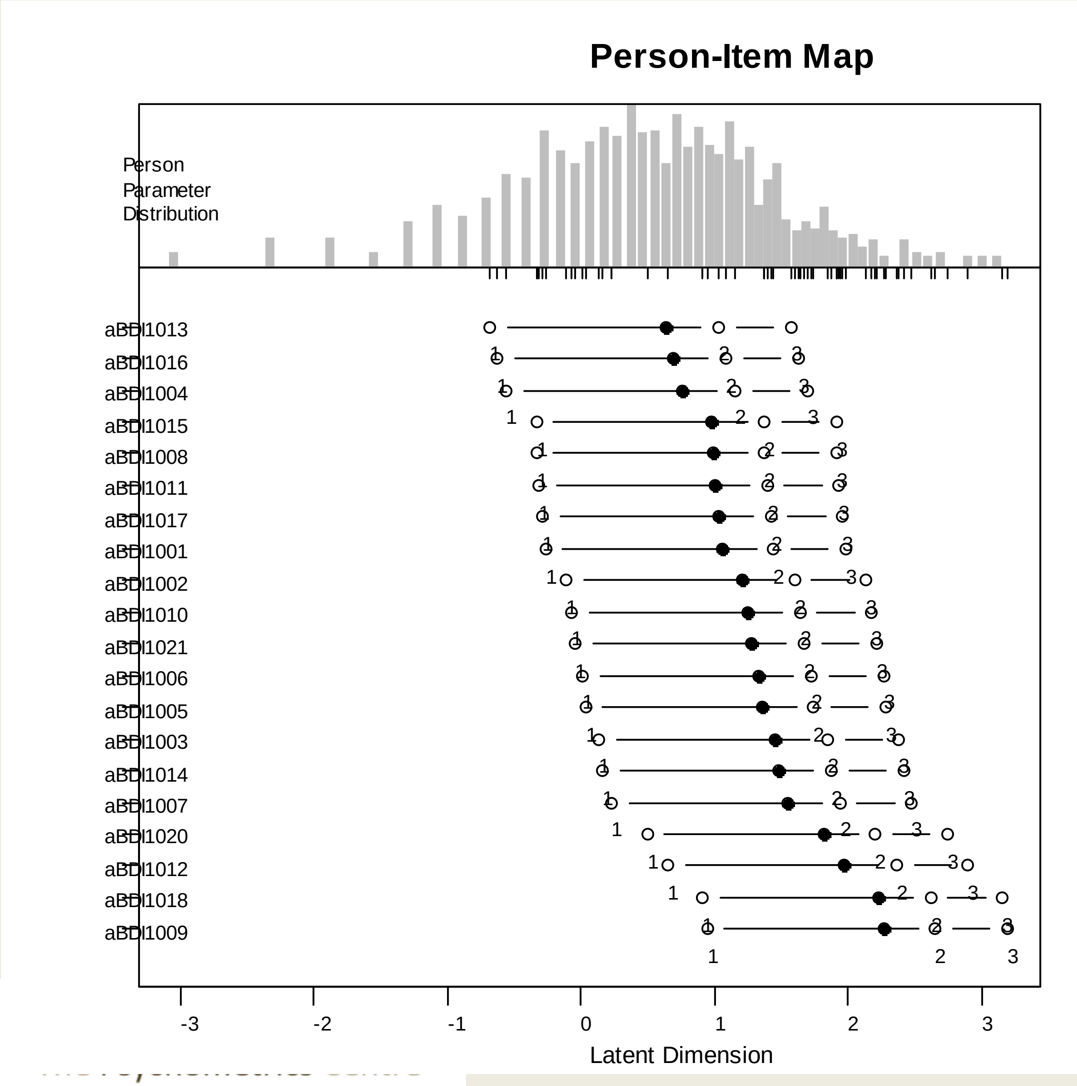
- starting with the restricted case of the RSM:

- The function "RSM" is used:

```
Result<-RSM(data, se=TRUE, sum0=TRUE)
```

Rating Scale Model

- circles: thresholds
- black dots: difficulty (comparable to item mean)



Rating Scale Model

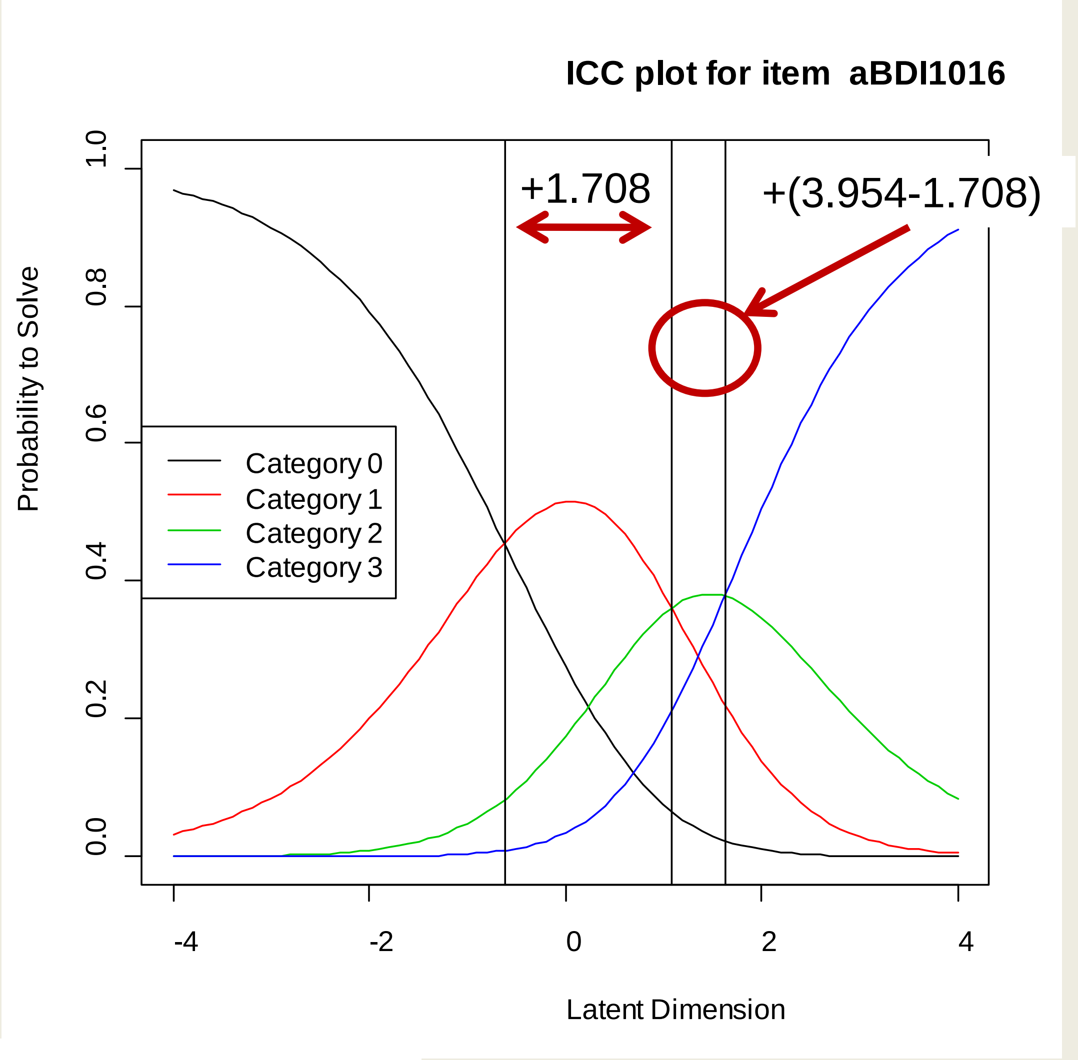
- the RSM imposes the exact same differences between category steps on every item
- in eRm estimated via
  - estimation of the first threshold
  - and estimation of difference parameters between first and second as well as first and third threshold

Rating Scale Model

- Category parameter 0/1: first threshold, estimated
- Category parameter 1/2: second threshold, 1.708
- Category parameter 2/3: third threshold, 3.954

Rating Scale Model

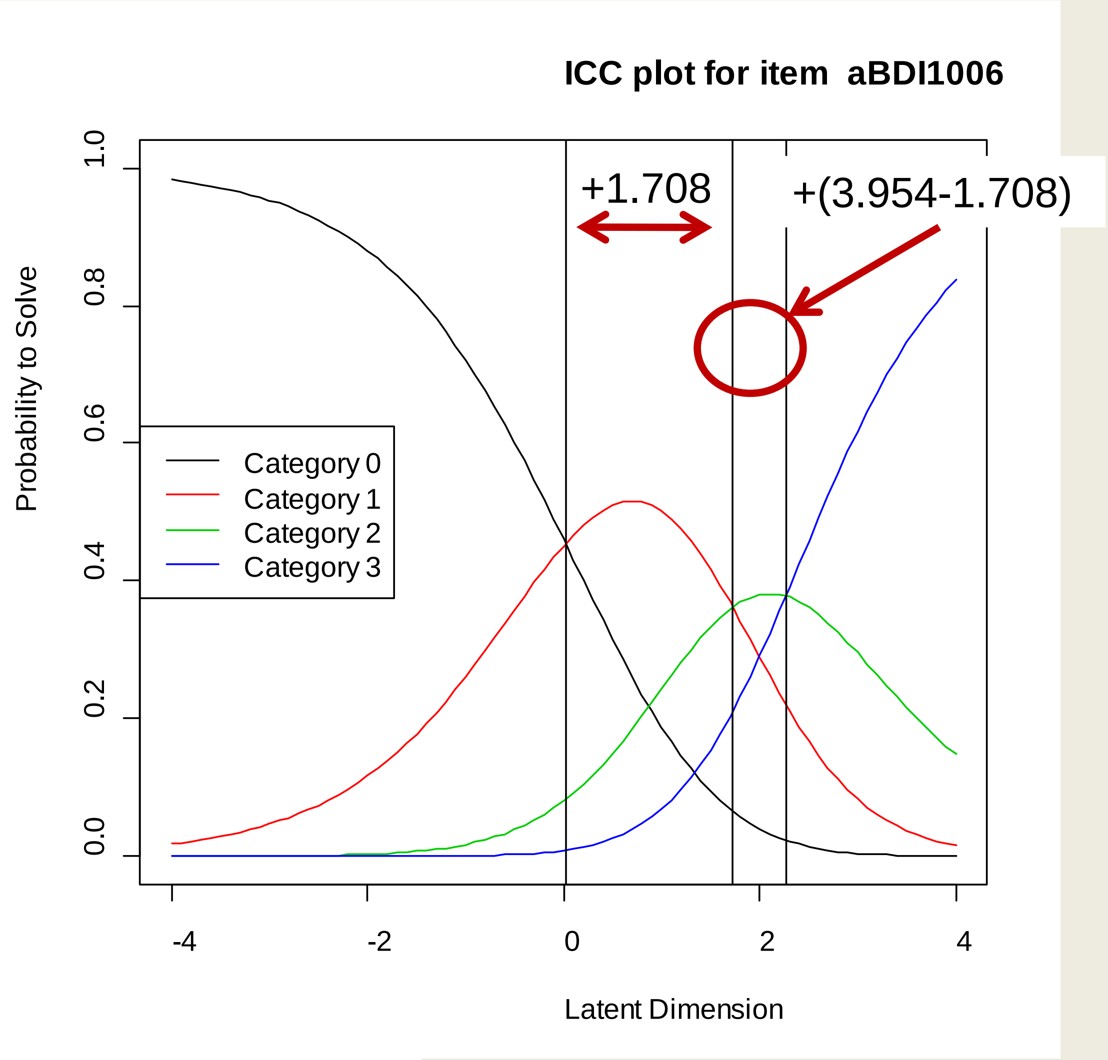
- Category parameter 0/1: first threshold, estimated; Item 16 (sleep disturbances): -.622



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Rating Scale Model

- Category parameter 0/1: first threshold, estimated; Item 06 (feeling / waiting to be punished): .018



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Differences RSM & PCM

- the major difference between these two models is
  - the PCM allows every item to have its own structure of category steps
  - whereas the RSM imposes the exact same differences between category steps on every item
  - (also models possible that use the same ratios etc)
  - AND every item can have its own number of categories

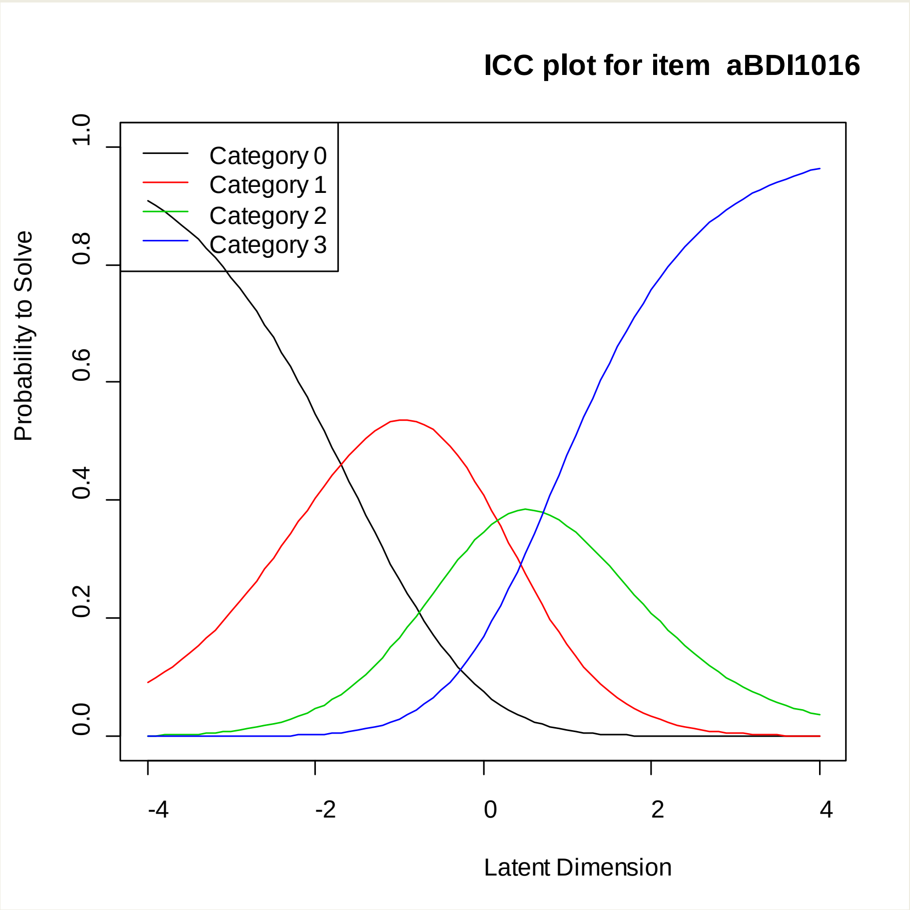
Differences RSM & PCM

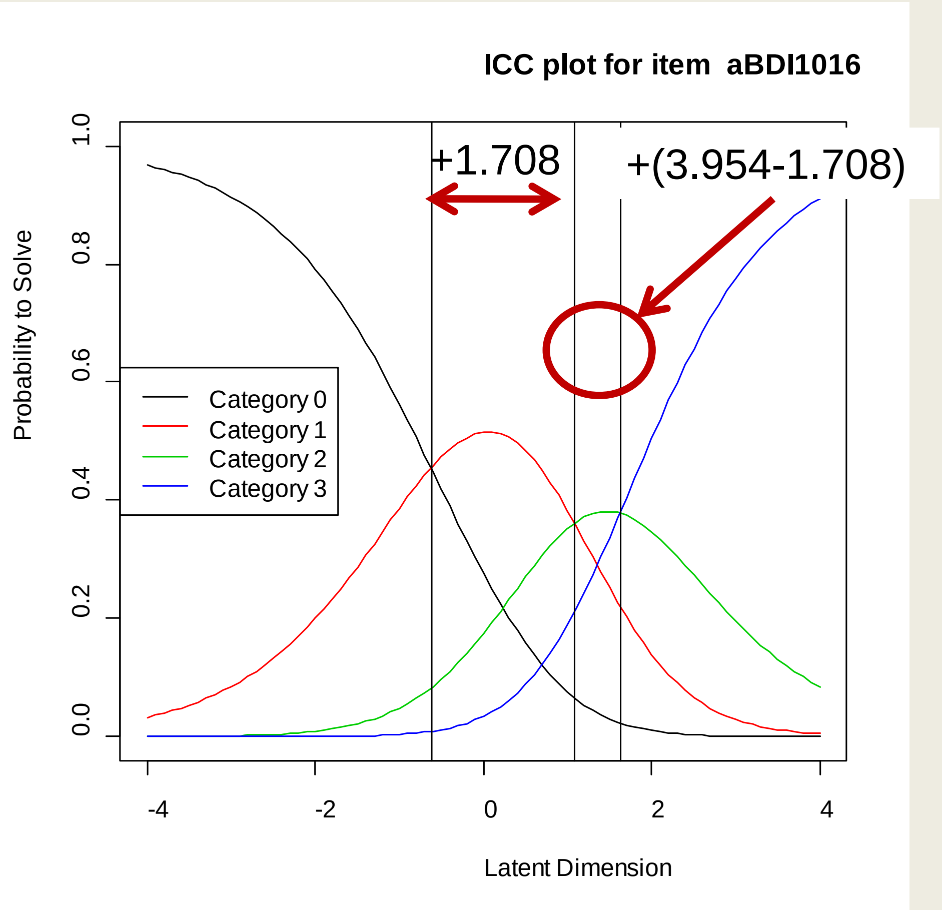
- the Partial Credit Model makes it possible that every item has its own pattern of thresholds
- in eRm estimated via
  - estimation of all thresholds of the items but one
  - (either parameterized that that have to sum to 0 or the first threshold is set to be 0)

Differences RSM & PCM

- Category parameter 0/1: first threshold, estimated
- Category parameter 1/2: second threshold, estimated
- Category parameter 2/3: third threshold, estimated

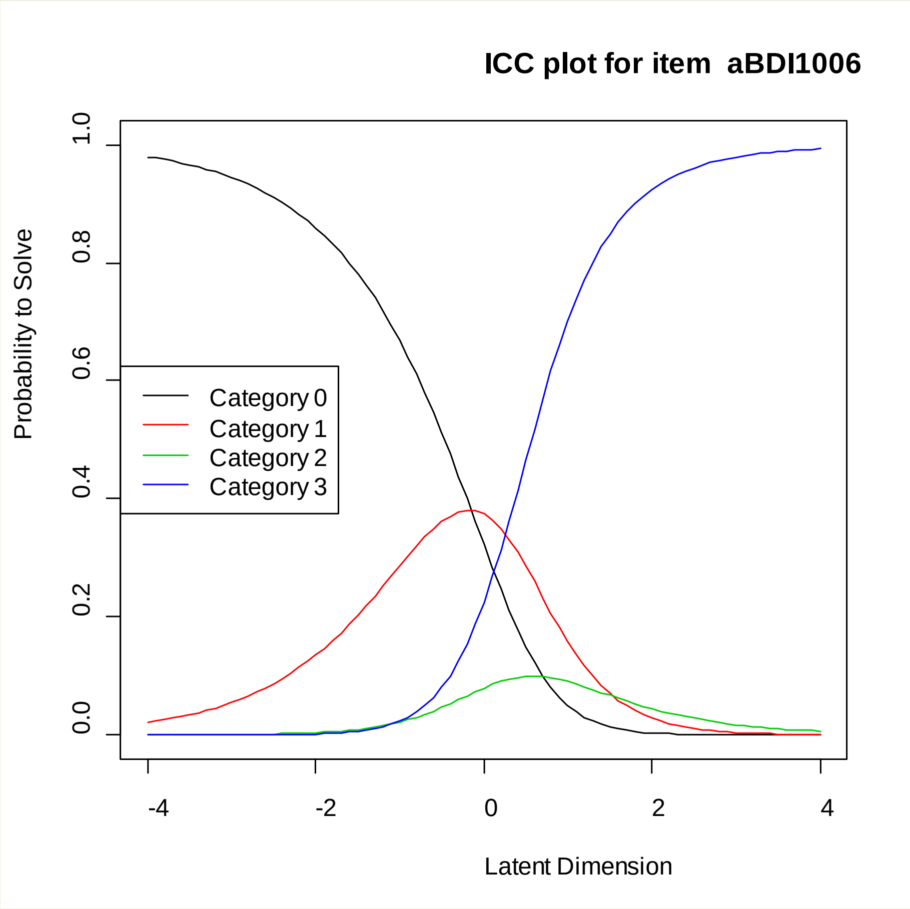
Differences RSM & PCM

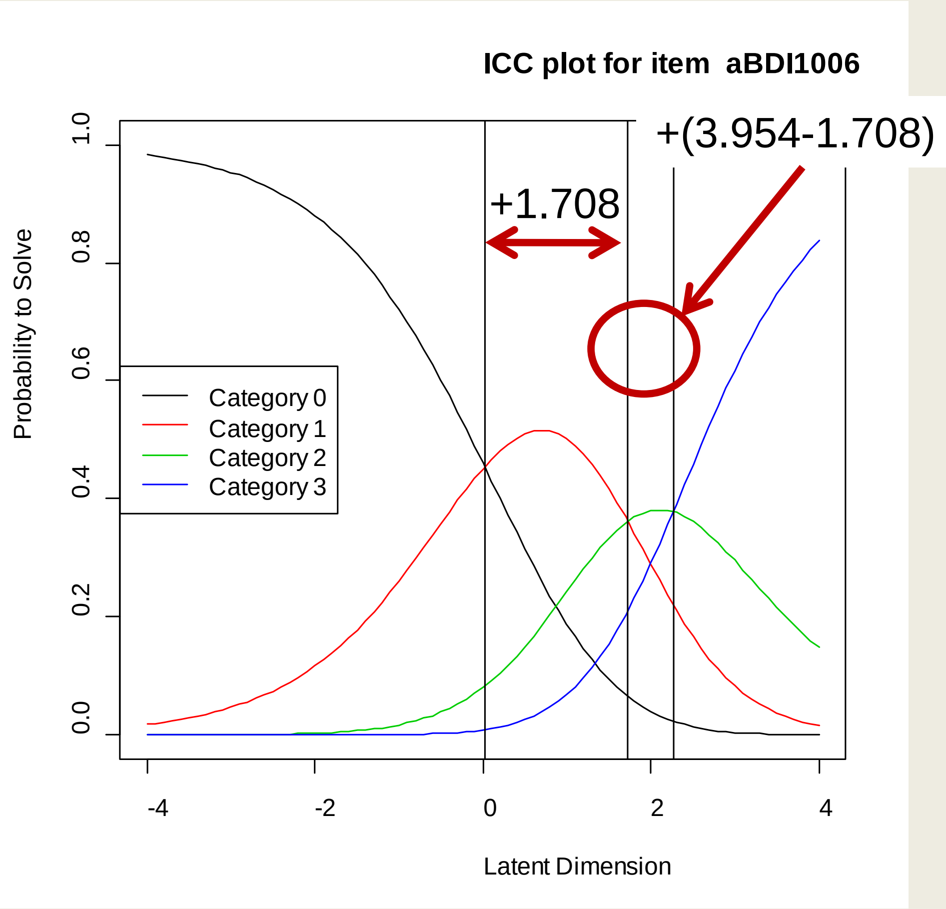




UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

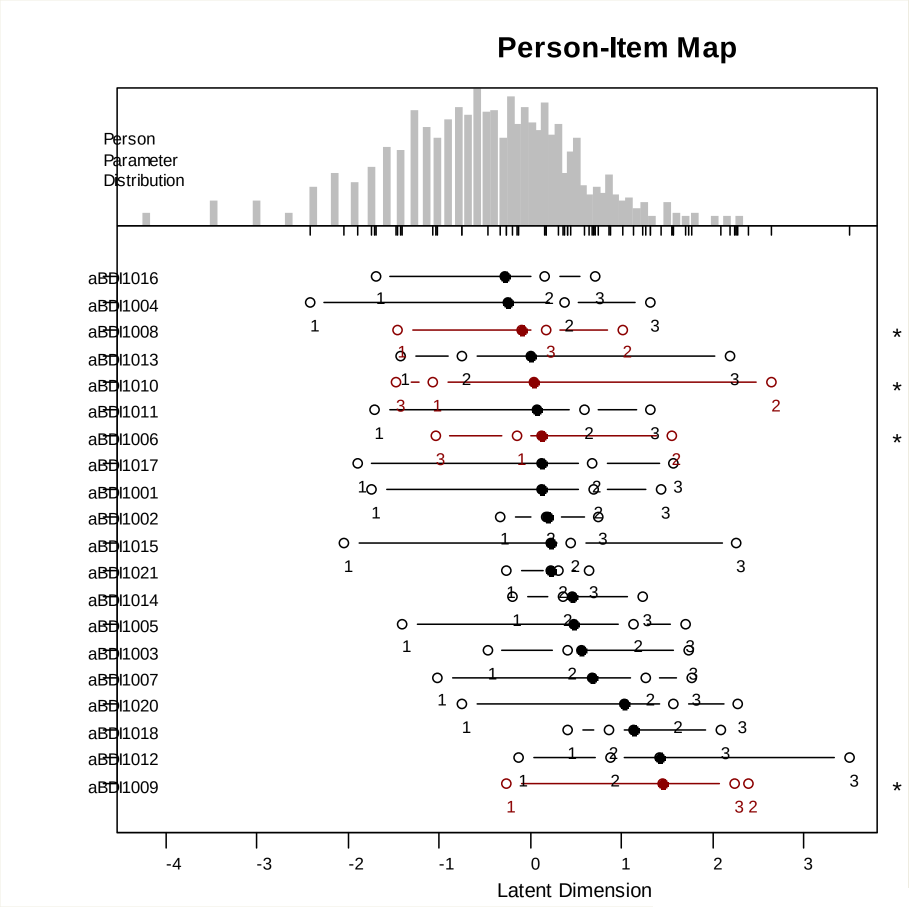
Differences RSM & PCM

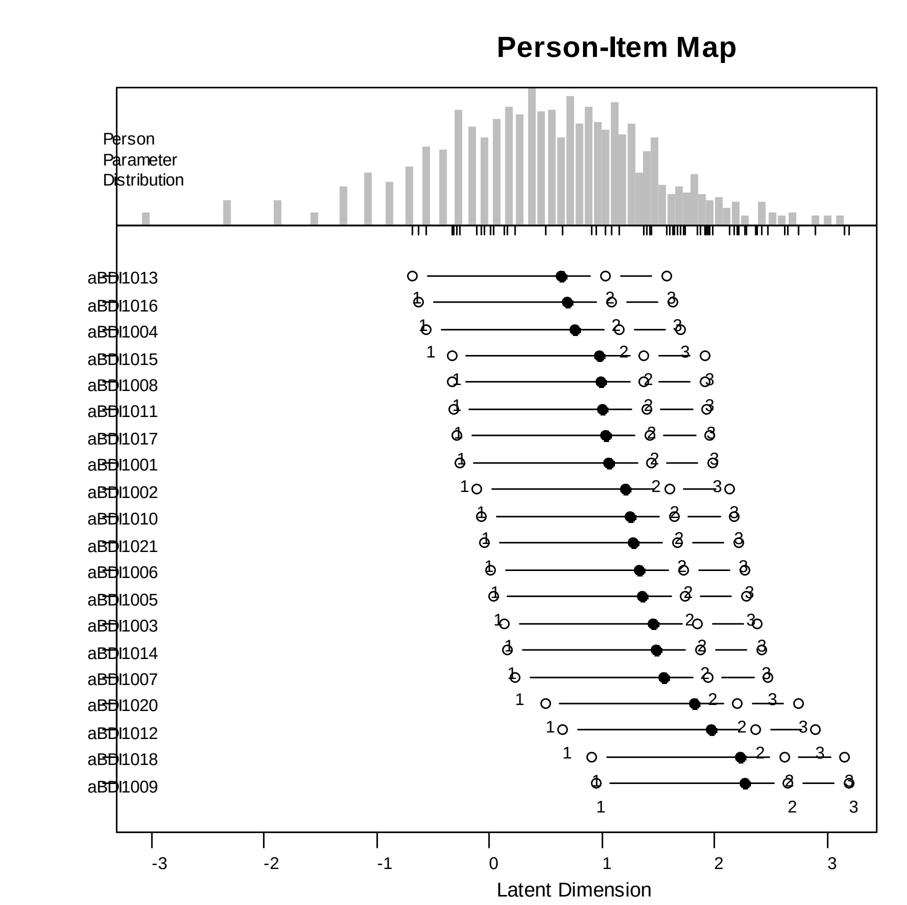




UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

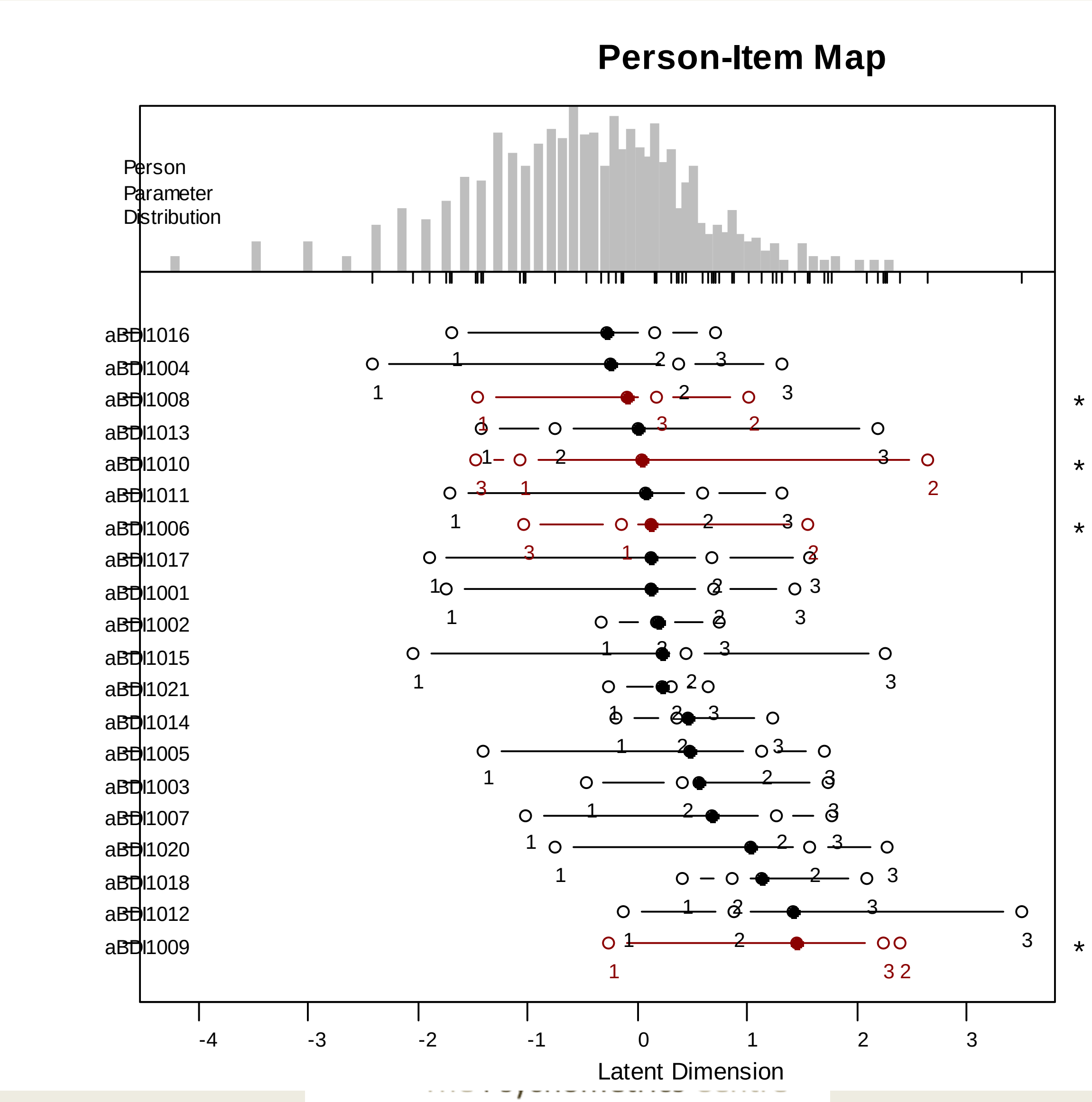
Differences RSM & PCM





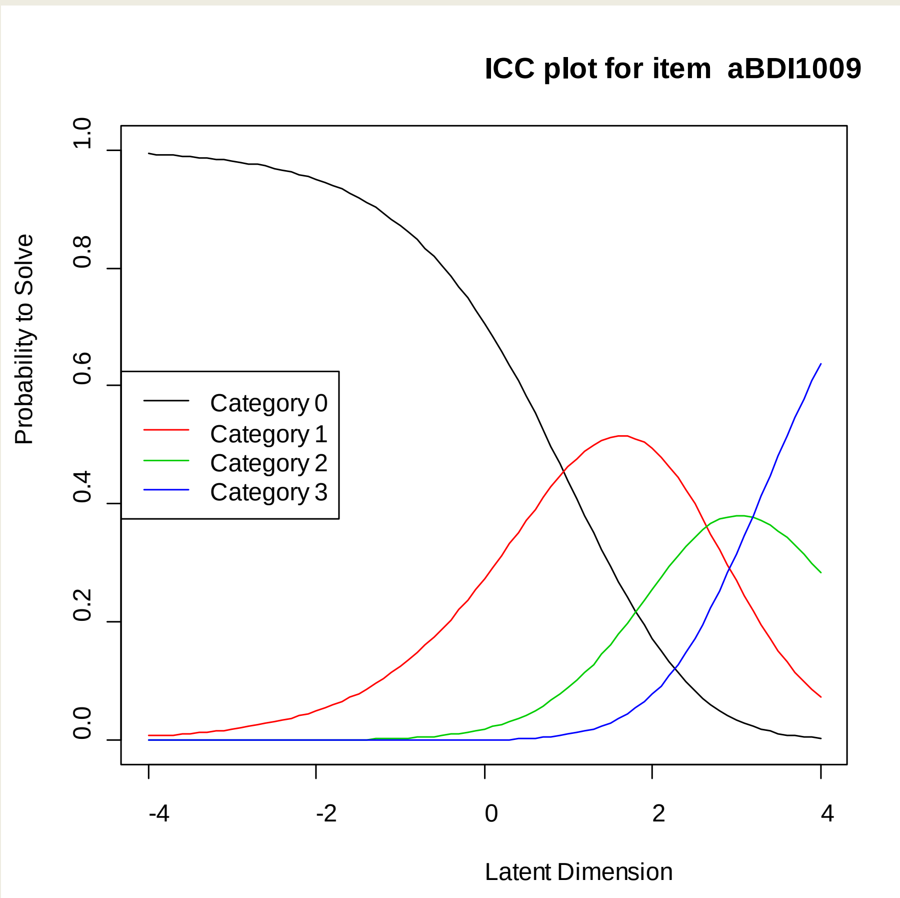
UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

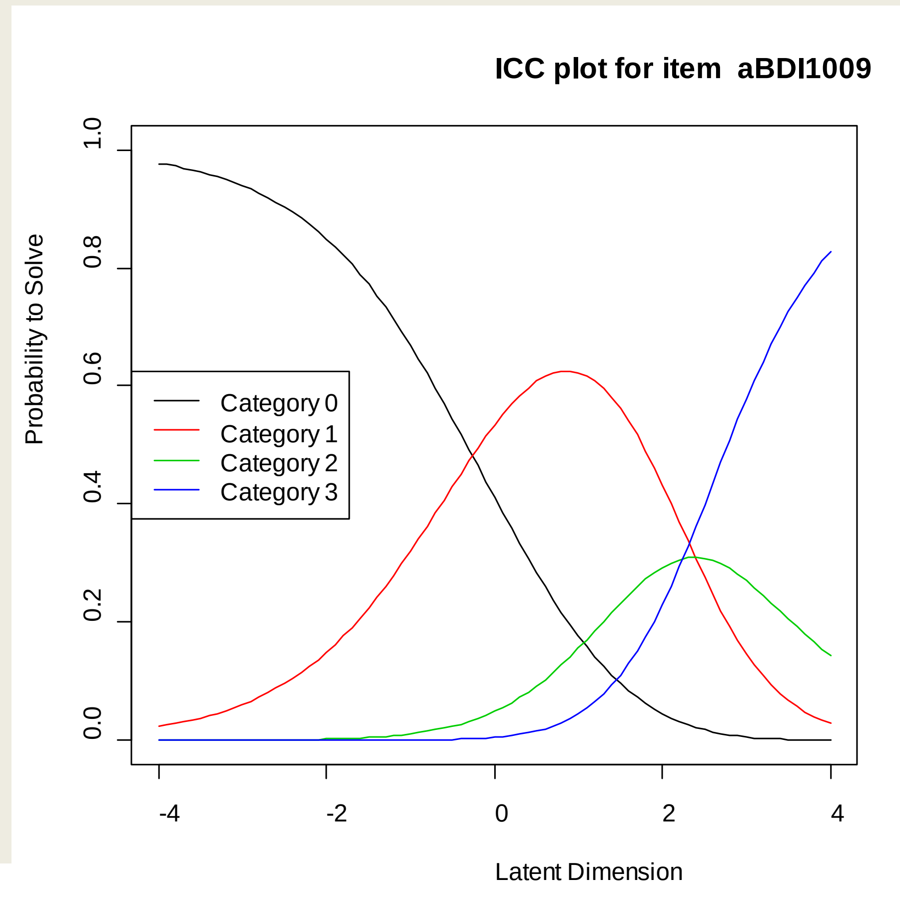
Differences RSM & PCM



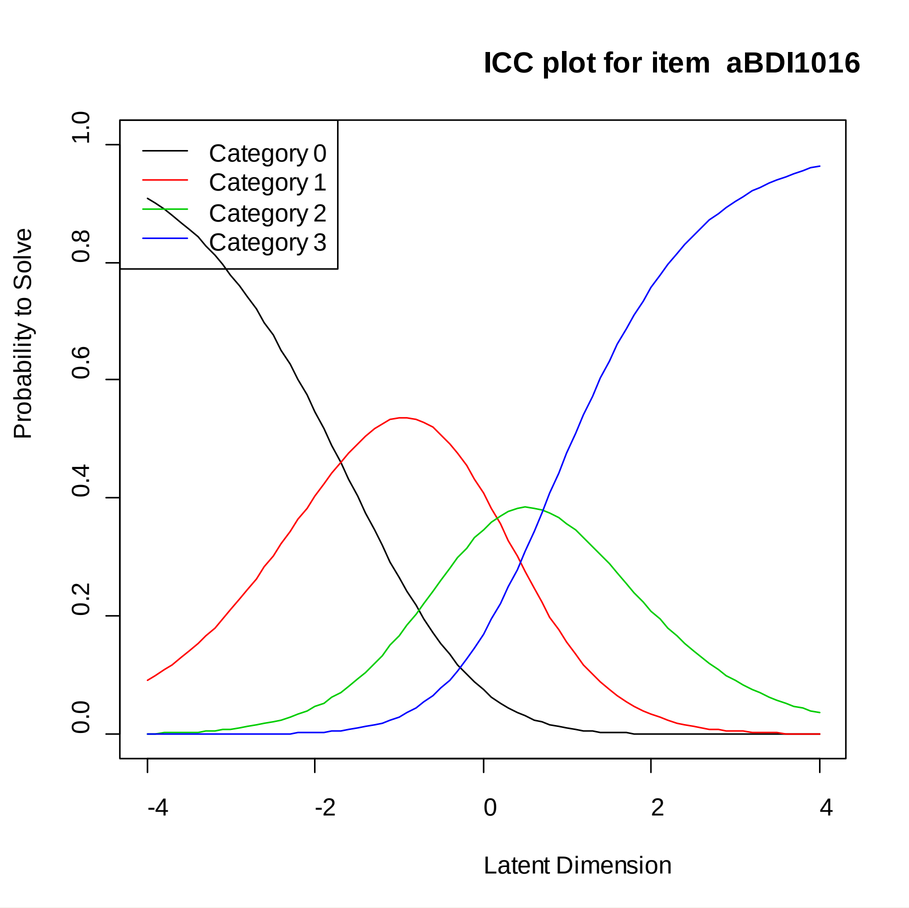
Differences RSM & PCM

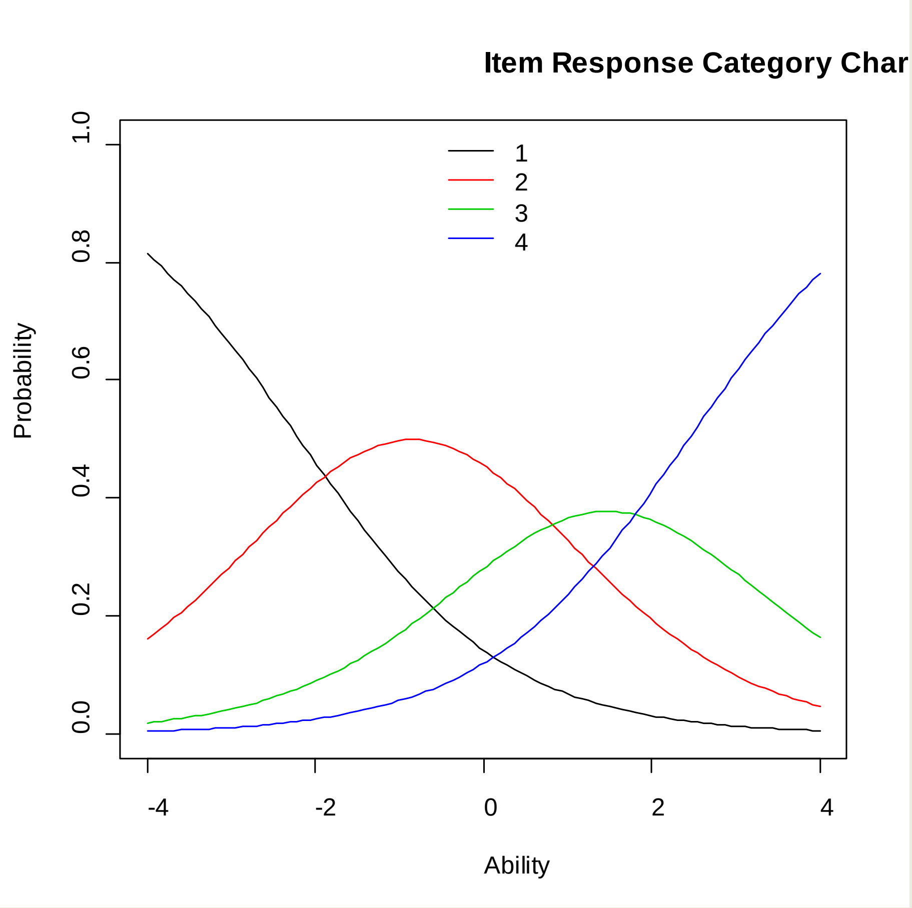
- BDI item 9, suicidal ideation



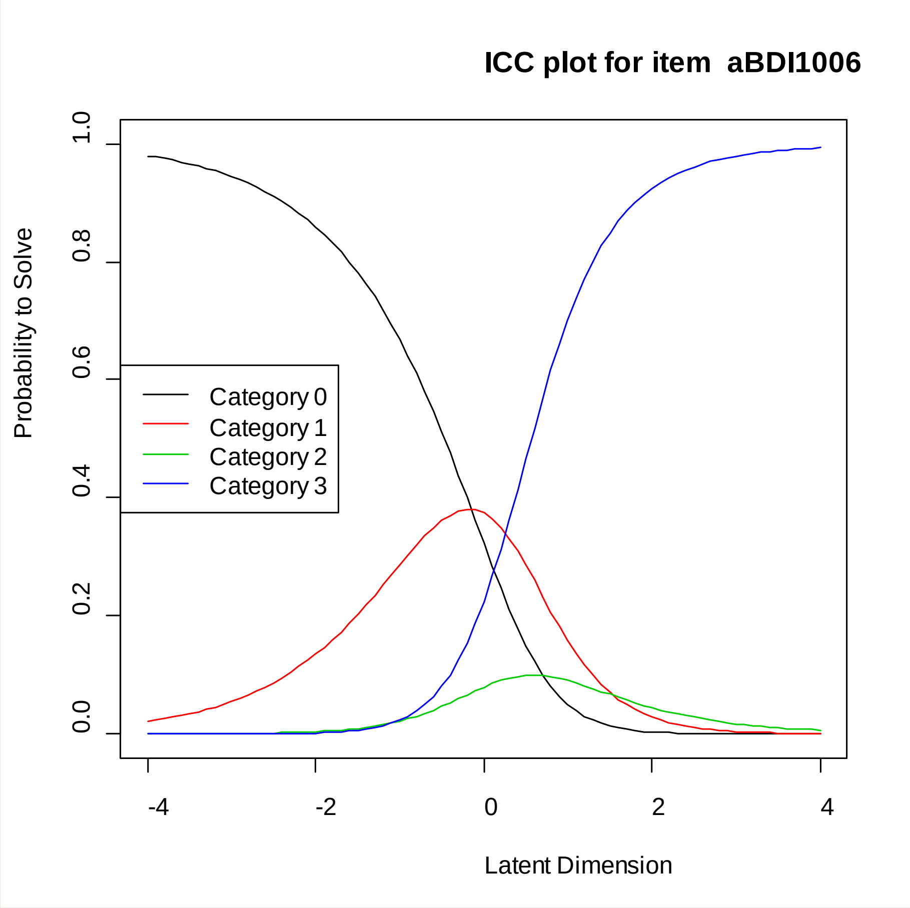


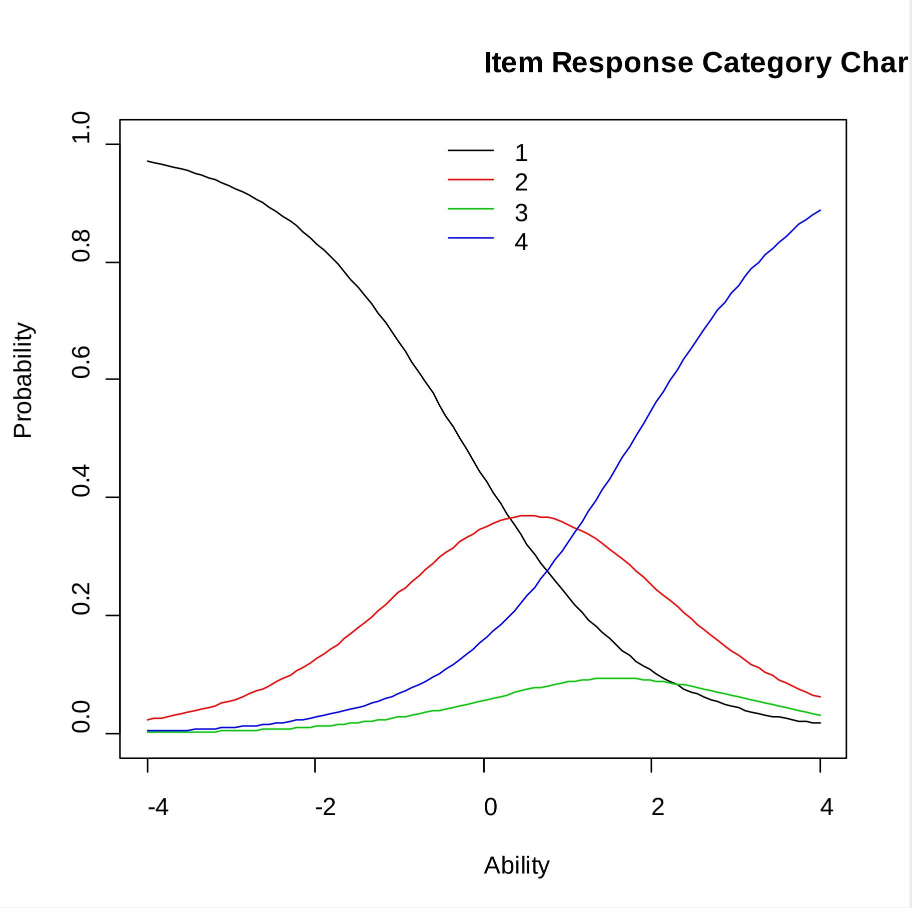
Differences PCM & ____



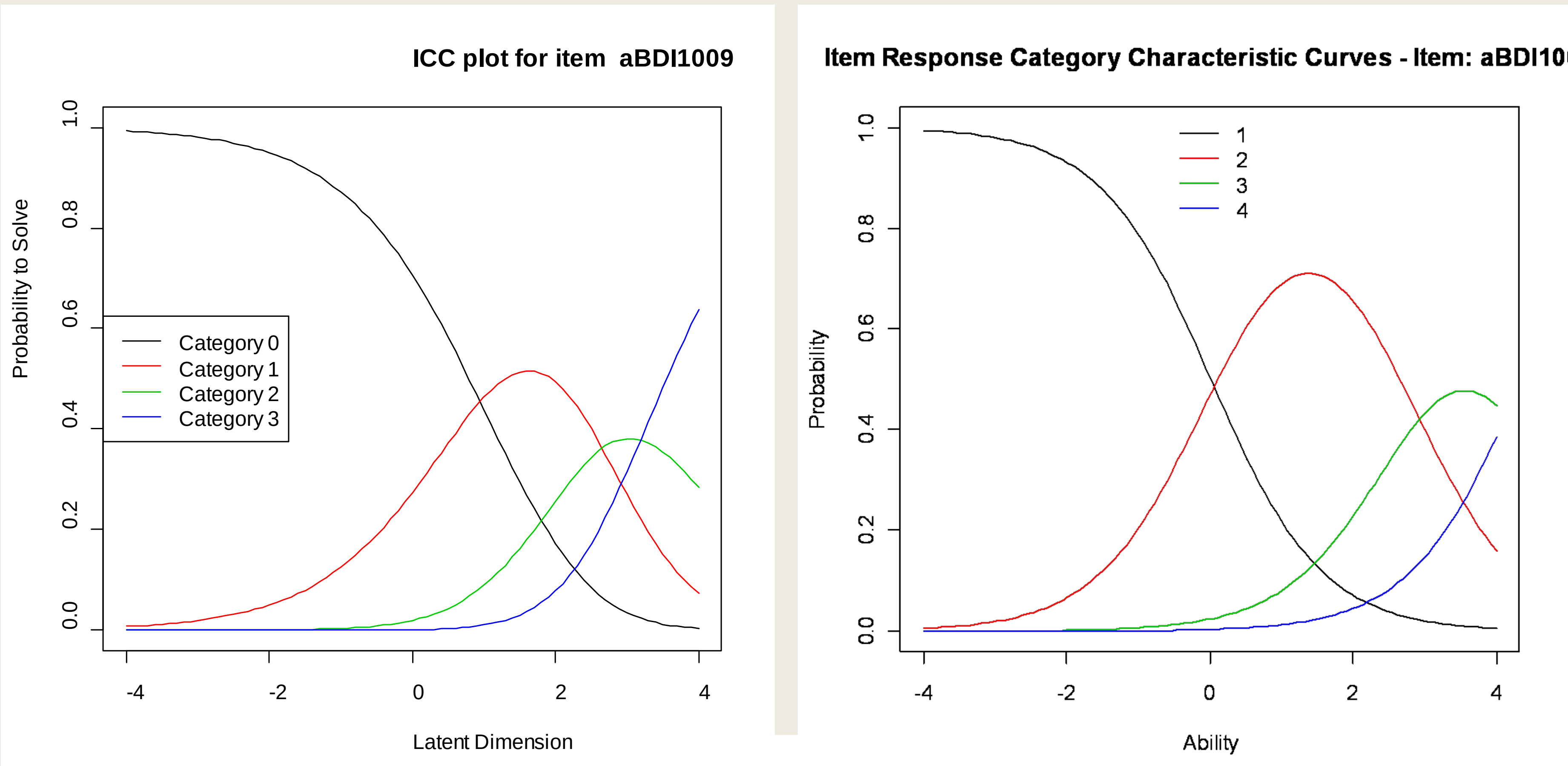


Differences PCM & ____





Differences PCM & ____



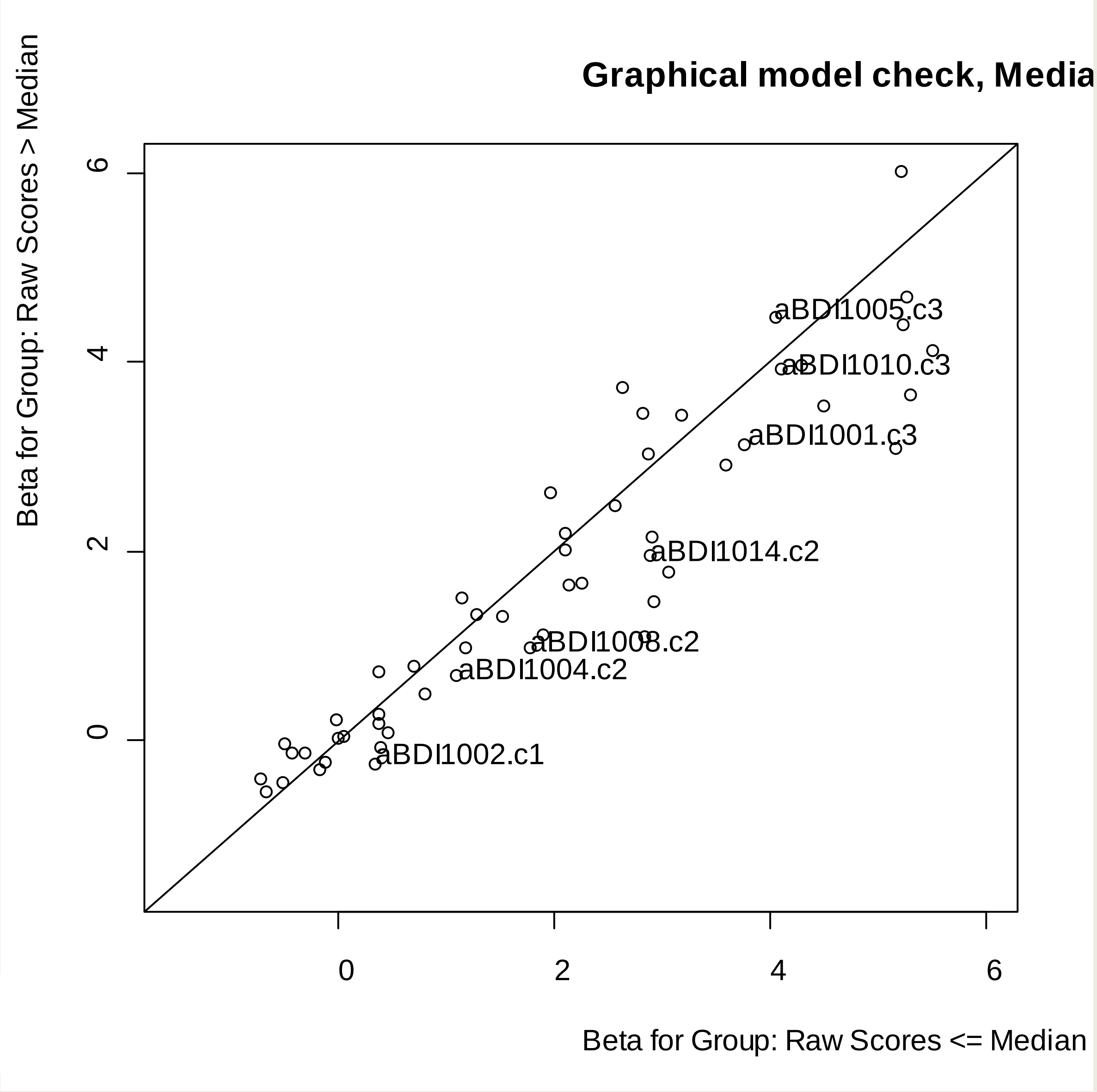
The Psychometrics Centre

Testing polytomous Rasch Models

- since in the estimation process for CML polytomous items are treated as if they were dichotomous items
- polytomous Rasch Models are testable in the same way as dichotomous Rasch Models

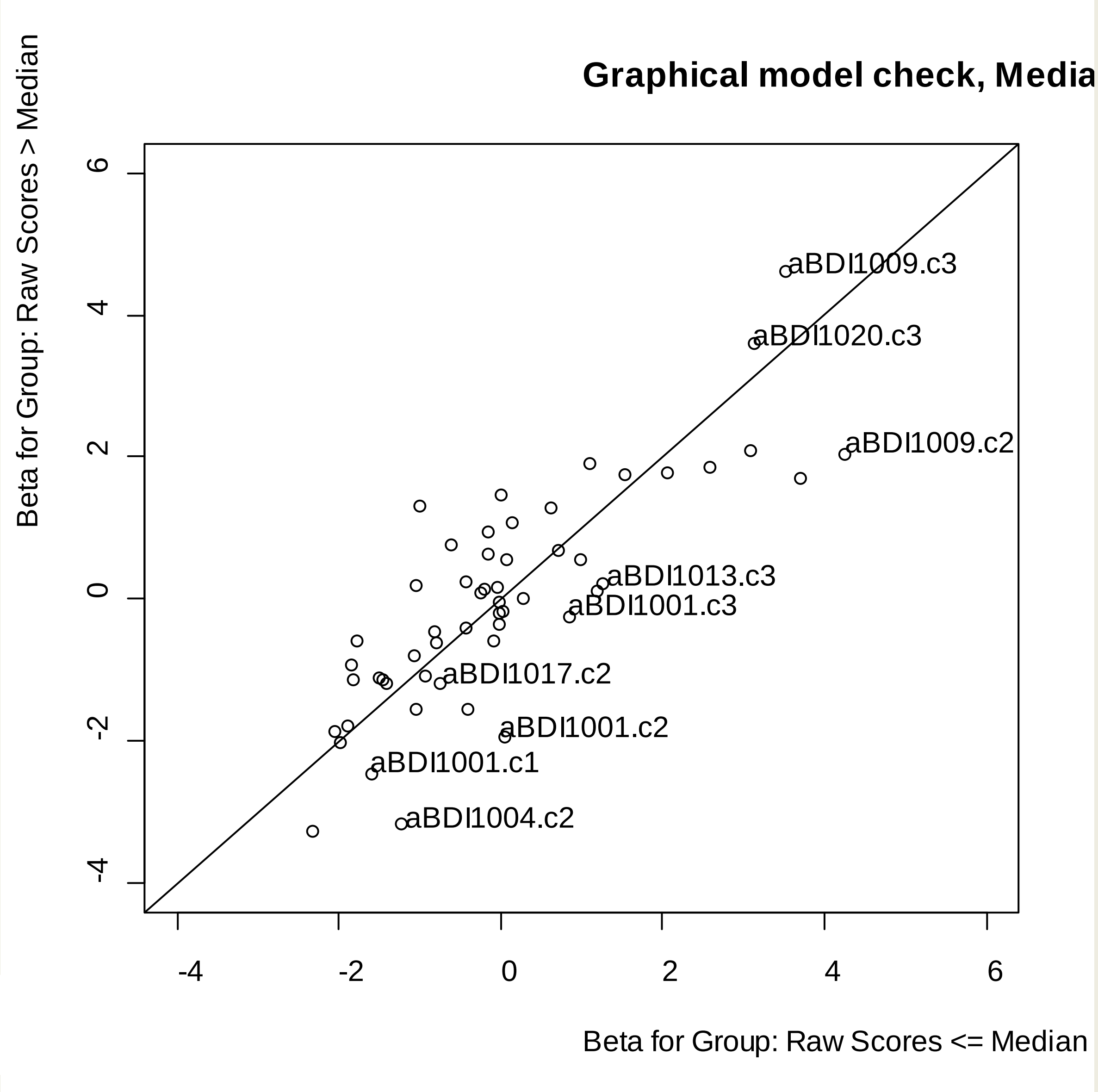
Testing polytomous Rasch Models

- Test with RSM
- p < .001



Testing polytomous Rasch Models

- Test with PCM
- p < .001



RSM vs PCM

- RSM needs substantially less parameters
- this was before the 2000s a substantial advantage

- today in my opinion no reason to use this model anymore
- (despite the case in which LR test between RSM and PCM shows no significant difference)

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…

Why Rasch?

- often critique: there are no data, that fit that model
- several responses are possible:
  - bad theories produce bad empirics
  - Rasch is a very simple model and reality is not simple (LLTM, LLRA, Mix-Rasch, Multidimensional- / Nominal-Rasch model,…)
  - BUT it is a model where in detail can be tested, whether it fits the data, or not

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.

University of Cambridge
The Psychometrics Centre

Choice of model must be pragmatic

- 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

Rasch as model of choice

- for many applications also models with more parameters might be able to reliably discriminate between different levels of a continuous latent trait

Rasch as model of choice

- but the Rasch Model it is the only test model that ensures specific objectivity and in which the local stochastic independence assumption is testable
- therefore, especially in high stakes testing situations the Rasch model proves to be extremely useful